

Vilniaus Pedagoginis Universitetas
Matematikos ir informatikos institutas

Romanas JANUŠKEVIČIUS

STATISTIKOS ĮVADAS

VILNIUS
2000

UDK 311(075.8)

Ja464

ISBN 9986-869-57-9

Romanas Januškevičius

Statistikos įvadas. Mokymo priemonė studentams

Tirdama masinių reiškinių kiekybinius aspektus, statistika siekia atskleisti tų reiškinių visumos bendrąsias savybes. Pirmoji vadovėlio dalis skirta intuityviajai statistikos daliai, elementariai dėstomi jos principai bei nagrinėjami paprasčiausi praktiniai pavyzdžiai. Tai draudimo ir kazino modeliai, matematinės vilties (vidurkio) fizikinė interpretacija, statistinių išvadų supratimas praktinių modelių kontekste. Antroji dalis yra teorinio pobūdžio, joje nagrinėjami matematinių statistikos metodų pradmenys. Knygos gale pateikiama terminų rodyklė, pagrindinių skirstinių reikšmių lentelės, vartojamų simbolių lentelė, paveikslų ir jų komentarų sąrašas.

Recenzavo dr. Marius Radavičius,

Matematikos ir informatikos instituto
Taikomosios statistikos skyriaus
vyresnysis mokslinis bendradarbis

Rinko ir maketavo Matematikos ir informatikos instituto
Kompiuterinės leidybos grupė

© Romanas Januškevičius

© Vilniaus pedagoginis universitetas

Tiražas 230 egz. 10,75 sp. l. Užsak. Nr. 41
Išleido Vilniaus pedagoginis universitetas,
Studentų g. 39, LT-2034 Vilnius

Spausdino VPU spaustuvė,
T. Ševčenkos 31, LT-2009 Vilnius
Kaina sutartinė

TURINYS

Pratarmė	6
ĮVADAS	9
Kas yra statistika?	9
I. ELEMENTARIOJI STATISTIKA IR JOS TAIKYMAI	10
1. Stochastinis eksperimentas. Statistinio stabilumo fenomenas	10
2. Elementariųjų įvykių erdvė. Atsitiktinis įvykis. Populiacijos ir imties sąvokos	12
3. Statistinių duomenų apdorojimas. Santykinis dažnis. Dažnių skirstinys	15
4. Tikimybinio mato įvedimas. Atsitiktinio įvykio tikimybė	20
5. Atsitiktiniai dydžiai. Tikimybės masės funkcija. Tikimybinė histograma	22
6. Populiacijos vidurkis. Didžiųjų skaičių dėsnis. Draudimo ir lošimo modeliai	25
7. Dispersija kaip sklaidos matas. Čebyšovo taisyklė duomenų analizei	29
8. Tikimybinio tankio kreivė kaip histogramų riba. Tankio savybės	32
9. Tolygusis skirstinys. Mediana, kvantilis, procentilis, kvartilis	36
10. Normalusis skirstinys ir jo tankio savybės	41
11. Statistinių išvadų supratimas praktinių modelių kontekste	45
12. Įvertinys ir įvertis. Empirinio vidurkio skirstinys. Gražintinis ėmimas	48
13. Nėgrąžintinis ėmimas. Korekcijos koeficientas. Vidurkio standartinė paklaida	52
14. Empirinė centrinės ribinės teoremos interpretacija	55
15. „Geriausio įvertinio“ problema. Taškinis ir intervalinis vertinimai. Nepaslinktasis įvertinys	59
16. Vidurkio pasiklivimo intervalas, kai dispersija yra žinoma. Paklaidos režis	61
17. Stjudento skirstinys. Vidurkio pasiklivimo intervalas, kai dispersija yra nežinoma	66
18. Chi kvadrato skirstinys. Dispersijos pasiklivimo intervalas	70
19. Hipotezių formulavimas. Hipotezių tikrinimo mechanizmo analizė	74

20. Sprendimo priėmimo taisyklė. Kritinė sritis	77
21. Didelių imčių normalumo kriterijus. Du klaidų tipai	80
22. Tiesinis sąryšis. Postūmio ir krypties koeficientai. Tiesinės koreliacijos koeficientas	82
23. Koreliacijos koeficiento apskaičiavimas ir jo savybės. Regresijos tiesės parinkimas	86
24. Prognozės lygtis. Mažiausių kvadratų metodas. Regresijos koeficientas	90
II. MATEMATINIAI STATISTIKOS METODAI	94
25. Diskrečioji elementariųjų įvykių erdvė ir jos mato savybės	94
26. Aibių algebra ir σ -algebra. Teorema apie minimalios σ -algebros egzistavimą	97
27. Borelio σ -algebra. Tikimybinis matas mačiojoje erdveje	100
28. Kolmogorovo aksiomų sistema. Tikimybinių erdvių	103
29. Atsitiktinis dydis kaip mačioji funkcija ir jo generuoti atsitiktiniai įvykiai	105
30. Tikimybių skirstinys. Skirstinio funkcija ir jos savybės	108
31. Veiksmas be praeities poveikio. Koši lygtis statistikoje	112
32. Empirinis matas. Teorema apie šių matų sekos konvergavimą pagal tikimybę	115
33. Variacinė eilutė, empiriniai skirstiniai. Glivenko – Kanteli teorema	118
34. Empirinių charakteristikų konvergavimas į jų teorinius analogus	121
35. Pagrindinis įvertinių konstravimo metodas. Suderintasis įvertinys	124
36. Silpno konvergavimo sąvoka. Empirinių charakteristikų asimptotinis normalumas	126
37. Momentų metodas kaip keitimo metodo realizacija	129
38. Paprasčiausias statistinių įvertinių palyginimas. Sistemingoji paklaida	131
39. Dominuojantysis matas. Minimizacijos lema	132
40. Didžiausio tikėtumo įvertinys ir jo apskaičiavimas konkrečiais atvejais	135
41. Paprastoji ir sudėtinė hipotezės. Hipotezių tipai. Praktinių ir teorinių rezultatų suderinamumas	138
42. Skirstinio atitikimas. Statistinis kriterijus (testas). Kritinė sritis	141
43. Chi kvadrato kriterijus (K. Pirsono suderinamumo testas)	144
44. Chi kvadrato kriterijaus praktinis taikymas	148

45. Kolmogorovo ir Smirnovo kriterijai suderinamumo ir vienodo pasiskirstymo hipotezėms tikrinti	153
PRIEDAS	
1 lentelė. Standartinio normaliojo skirstinio reikšmės	157
2 lentelė. Standartinio normaliojo skirstinio uodegų reikšmės	158
3 lentelė. Chi kvadrato skirstinių reikšmės	159
4 lentelė. Stjudento skirstinių reikšmės	160
Naudota literatūra	161
Paveikslų ir jų komentarų sąrašas	162
Rodyklė	165
Vartojamų simbolių lentelė	171

PRATARMĖ

1. Pirmas akmuo į autoriaus daržą. „Jei žmogus, pratęs dirbti protinį darbą, susiduria su praktinio veiksmo būtinybe, jam rūpi viską komplikuoti, daryti pernelyg gudriai“. Tai citata iš XX amžiaus aukso fondo serijos knygos, priklausanti Donnai Tartt.

Panašiai atsitiko su šių eilučių autoriumi. Pagal JAV ir Vokietijos matematikos referatyvinių žurnalų klasifikatorių, abi mano disertacijos priskirtinos temai 62E10 „Charakterizacija ir struktūrų teorija“. Gi visa ši tema priskirta ne tikimybių teorijai, o statistikai, kurios strategiją formuoja praktiniai poreikiai. Tačiau autoriaus darbai buvo grynai teorinio pobūdžio.

Tai ir sąlygojo statistikos kurso, vieną semestrą skaitomo VPU matematikos magistrantams, pirmojo varianto turinį. Šis kursas prasidėdavo nuo σ -algebros apibrėžimo ir atsitiktinio dydžio kaip mažiosios funkcijos analizės. Tokią teorinę kurso orientaciją grindžiau tuo, kad, mano supratimu, būsimajam pedagogui yra svarbiausia išmokti logiškai mąstyti pačiam, o po to išmokyti mąstyti ir mokinius.

Ir nors statistikos egzamino rezultatai buvo neblogi, su viena metodikos doktorante iškilo konfliktėlis. Doktorantė tvirtino, jog jai visai nereikia žinoti nei kas yra σ -algebra, nei kas yra mažioji funkcija. Ji norėjo žinoti, kaip teisingai rinkti duomenis ir moksliškai juos analizuoti, o šios analizės pagrindu teikti reiškinį vystymosi prognozes, ir visa tai būtų jos disertacijos pagrindas.

Tai buvo pirmas akmuo, mestas į mano matematinį darželį, subrandintą A.A. Borovkovo knygos „Matematinė statistika“ pavyzdžiu (naudotų knygų sąrašė – [1] pozicija).

2. Ar gali būti matematika per daug matematizuota? Praleisiu tą dalį, kurioje galėtų būti dėstoma, kaip tas matematinis darželis buvo ravimas. Tačiau paminėsiu, kad dar vienas akmuo tenai įkrito iš anoniminio Lietuvos mokslo ir studijų fondo eksperto pusės. Mat, savo pirmąjį paskaitų konspekto variantą pateikiau kaip pagrindą vadovėliui parašyti į fondo vadovėlių konkursą.

Pačios recenzijos man nedavė, o tik perskaitė, todėl galiu cituoti tik iš atminties: *vadovėlis pernelyg matematizuotas, jis būsimiesiems pedagogams net žalingas, nes pateikiamą medžiagą jie negali suprasti, o tik iškalti.*

„Tai bent! – pamaniau. – Ar gali būti sviestas pernelyg sviestuotas? Be to, egzamino rezultatų vidurkis – 7,3. Buvo atveju, kai magistrantas prašė

perlaikyti egzaminą, norėdamas 9 ištaisyti į 10. Ir tik vieną kartą magistrantė, gavusi klausimą apie hipotezių tikrinimą, prisipažino, kad atsakymo nesuprato, todėl iškalė“.

Fondas atmetė rankraštį, tačiau pasiūlė dalyvauti konkurse kitais metais. Buvo ankstyvas pavasaris. Buvo pats laikas pamastyti apie visiškai naują darželį.

3. Ar galima statistiką dėstyti elementariai? Bent jau be integralinio ir diferencialinio aparato, be σ -algebros apibrėžimo ir mažiosios funkcijos sąvokos? Ar galima sudėtingą medžiagą taip paprastai išaiškinti, kad šios medžiagos turinys ne tik nenukentėtų, bet atsiskleistų visoje savo pilnumoje?

Susipažinęs su R.A. Johnson, G.K. Bhattacharyya vadovėliu „Statistics. Principles and methods“ (naudotų knygų sąrašė – [5] pozicija) bei R. Khazanie vadovėliu „Statistics in a world of applications“ ([6] pozicija), išlaikiusiais ne vieną ir ne du leidimus, supratau, kad atsakymas į visus iškeltus klausimus yra vienas: „Taip!“

Detaliai analizuojami praktiniai pavyzdžiai leido visiškai kitaip pažvelgti ir į senus klasikinius vadovėlius – pavyzdžiui, į H. Cramer „Mathematical methods of statistics“ (žr. [7]). Savo matematine dvasia jam artimas naujas K.J. Hastings vadovėlis „Probability and statistics“ (žr. [3]).

Visi šie vadovėliai turi vieną trūkumą – juose pateikiamą medžiagą neišdėstyti nei per vieną, nei per du semestrus.

Taip gimė mintis paskaitų konspektą sudaryti iš dviejų dalių. Pirma dalis „Elementarioji statistika ir jos taikymai“ – praktinio pobūdžio. Joje nėra nei vieno integralo ženklo, nėra nei σ -algebros, nei mažiosios funkcijos sąvokų. Antroji dalis „Matematiniai statistikos metodai“ yra teorinio pobūdžio. Tai matematinių statistikos metodų pradmenys.

4. Kodėl Levas Tolstojus rašė aritmetikos vadovėlį? Beje, ne tik rašė, bet ir 1874 metais pats išleido. Ir jau kaip matematikos mokytojas mokė: „Tam, kad turėtum įsitikinimą, jog atneši naudą, reikia turėti vieną bruožą. Šis bruožas yra meilė.“

Panašus teiginys yra ir 1994-aisiais Vilniuje išleistose „Lietuvos bendrojo lavinimo mokyklos Bendrosiose programose“. Cituoju: „Mokantis matematikos, ugdomas sugebėjimas analizuoti, argumentuoti, klasifikuoti, kelti hipotezes, jas paneigti arba įrodyti.

Matematika gali žymiai prisidėti ne tik prie bendro asmenybės ugdymo,

bet ir prie charakterio bei dorinių nuostatų formavimo.

Matematika prisideda ir prie estetinio pasaulio suvokimo ugdymo“.

Pasiūlymas studentams – kol dar nevēlu, mokytis mąstyti. Kas gali išmokyti mąstyti, jei ne matematikai? Išmokę mąstyti, pajusime, kas yra grožis ir kas yra meilė.

Mąstymas – vienintelis dalykas, kurio nerasi enciklopedijose ir žinyuose.

5. Kas svarbiau – „Matematikos terminų žodynas“ ar „Lietuvos standartas“? Kuris terminas turi būti vartojamas – žodyno [10] „imties didumas“ ar standarto [8] „imties dydis“? Žodyno „pasikliautinis intervalas“ ar standarto „pasiklovimo intervalas“? Žodyno „įvertis“ ar standarto „įvertinys“ (angl. *estimator*) bei „įvertis“ (angl. *estimate*)? Išvardytais atvejais buvo pasirinktas antrasis, t.y. standarto terminas. Aplamai, stengtasi, kad studentas gerbtų valstybinį standartą ir mokėtų juo naudotis. Pavyzdžių, kuriuose naudojama Lietuvos statistinė medžiaga, duomenys yra paimti iš Lietuvos statistikos metraščio [9].

Autorius dėkoja profesoriui P. Survilai už pasiūlymą skaityti matematinės statistikos kursą VPU matematikos magistrantams bei pasiūlymą parašyti ir išleisti šį vadovėlį.

Romanas Januškevičius,

Matematikos ir informatikos instituto
Tikimybių teorijos skyriaus
vyriausiasis mokslinis bendradarbis,
Vilniaus pedagoginio universiteto
Algebros ir skaičių teorijos katedros
profesorius

IVADAS

KAS YRA STATISTIKA?

Žodis *statistika* yra kilęs iš lotyniško žodžio *status*, reiškiančio *būseną, padėtį*. Jau ne vieną šimtmetį žmonės domina duomenys apie demografinius ir aplinkos procesus, šalių geografiniai ir gamtiniai ištekliai, gamybos ir prekybos apimtys ir t.t.

Ilgainiui susidarė nuomonė, kad statistika – tai išimtinai skaičių rinkimas ir diagramų bei grafikų sudarymas. Toks įsitikinimas daugeliui ir šiandien atrodo teisingas gal dar ir todėl, kad žiniasklaidoje mes nuolat sutinkame duomenis ir pranešimus apie ekonominį gyventojų užimtumą ir aktyvumą, apie pagrindinius mokesčius į Lietuvos biudžetą, gyventojų sveikatą ir socialinę apsaugą, kainas ir kainų indeksus ir t.t.

Žinoma, duomenų rinkimas, diagramų bei grafikų sudarymas ir šiandien, t.y. XX amžiaus pabaigoje, yra minorinis statistikos aspektas, reikalaujantis nemažai rutininio darbo. Tad iškyla natūralus klausimas – koks gi yra statistikos kaip mokslinės disciplinos vaidmuo?

Atsakymas į šį klausimą buvo rastas dar XVII amžiuje. Mat jau tada statistika turėjo dvi kryptis – aprašomąją (duomenų rinkimas, diagramų bei grafikų sudarymas ir pan.) ir dedukcinę matematinę.

Tirdama masinių reiškinių kiekybinius aspektus, statistika siekia atskleisti tų reiškinių visumos bendrąsias savybes.

Štai keli tipiški klausimai, kuriuos nagrinėja statistika. Kokios rūšies ir kiek duomenų reikia surinkti? Kaip reiktų tuos duomenis rinkti ir kaip juos interpretuoti? Kaip analizuoti duomenis ir iš jų analizės daryti teisingas išvadas?

Pastaruoju metu terminu *statistika* priimta vadinti dvi disciplinas – tikimybių teoriją ir matematinę statistiką. Tikimybių teorijos vystymosi dėka XX amžiaus pradžioje ir ypač per pastaruosius 50-60 metų statistiniai metodai buvo iš esmės išvystyti ir buvo galima kalbėti apie naujos disciplinos – *matematinės statistikos* – atsiradimą ir spartų vystymąsi.

Pirmoji šio vadovėlio dalis skirta intuityviajai statistikos daliai, elementariai dėstomi jos principai bei nagrinėjami paprasčiausi praktiniai pavyzdžiai. Antroji dalis – tai įvadas į matematinius statistikos metodus.

I. ELEMENTARIOJI STATISTIKA IR JOS TAIKYMAI

1. STOCHASTINIS EKSPERIMENTAS. STATISTINIO STABILUMO FENOMENAS

1. Įvairiausiuose praktinių ir teorinių tyrimų srityse dažnai sutinkami atvejai, kai eksperimentai arba stebėjimai tose pačiose sąlygose gali būti pakartoti didelį skaičių kartų. Kiekvieną tokį kartą mes koncentruosime savo dėmesį į eksperimento (arba stebėjimo) rezultatą, išreikštą skaitine charakteristika.

Eksperimentus, kuriuos, esant vienodoms sąlygoms, galima pakartoti daugelį kartų ir kurių konkretaus rezultato negalima nuspėti iš anksto, vadiname stochastiniais (tikimybiniais) eksperimentais.

2. Paprasčiausias stochastinio eksperimento pavyzdys – monetos metimas. Jo rezultatas aiškus – pasirodys arba herbas, arba skaičius. Jei herbo pasirodymą pažymėsime 1, o skaičiaus – 0, tai šio eksperimento rezultatus išreikšime skaičiais. Santuokų ir ištuokų registravimas metų bėgyje valstybėje, apskrityje ar mieste – jau sudėtingesnio stochastinio eksperimento pavyzdys. Suskaičiavus santuokas ir ištuokas Lietuvoje, gauta, kad jų, pavyzdžiui, 1990-aisiais yra atitinkamai 36310 ir 12747, o po septynerių metų, t.y. 1997-aisiais, – jau tik 18796 ir 11371. Šio eksperimento rezultatų dar po septynerių metų, t.y. 2004-aisiais, aišku, negalima tiksliai nuspėti iš anksto, tačiau tendencijų prognozė yra svarbi.

3. Kiekviena stochastinių eksperimentų sekos rezultatų suvestinė sudaro statistinių duomenų aibę.

Pagrindinis statistikos objektas yra galimybių gauti patikimas išvadas statistinių duomenų pagrindu tyrimas ir metodų, leidžiančių pateikti šias išvadas, išvystymas.

4. H. Krameras¹ 1946-aisiais metais rašė, kad „turbūt neįmanoma pateikti tikslų apibrėžimą to, ką nusako žodis *atsitiktinis*“, ir šio žodžio prasmę aiškina pavyzdžiu.

Jei daug kartų mesti paprasčiausią monetą, tai, net labai stengiantis išsaugoti eksperimento sąlygas nepakitusiomis, mes vis tiek nepajėgsime pasakyti, ar konkretaus metimo atveju iškris herbas, ar skaičius. Net jei mes sukonstruosime mašiną, kiekvieną kartą metančią monetą visiškai vienodai, tai visgi neįtikėtina,

¹ Harald Cramér – švedų matematikas.

kad mums pavyktų tiksliai nuspėti konkretaus metimo rezultata.

5. Tai galima paaiškinti tuo, kad nežymūs stebimo objekto pradinių sąlygų pakitimai, kuriuos neįmanoma užfiksuoti mūsų „instrumentais“, gali žymiai įtakoti galutinį rezultatą. Ir net jei tokie „instrumentai“ būtų, stebimas reiškinys gali būti toks sudėtingas, kad jo parametrų praktinis apskaičiavimas ar net teorinis aprašymas gali būti neįmanomas.

Netrukus įsitikinsime, kad visuose tokiuose atvejuose tarp visų stochastinio eksperimento pradinių sąlygų pakitimų ir jo rezultatų svyravimų galima aptikti vieną dėsningumą, kuris ir yra matematinės statistikos pagrindas. Kalbėsime apie *statistinio stabilumo fenomeną*.

6. Taigi, atliekant konkretų stochastinį eksperimentą yra neįmanoma nusakyti jo rezultata. Tačiau nuo konkretaus eksperimento perėjus prie stochastinių eksperimentų sekos, nagrinėjama situacija radikaliai pasikeičia.

Pastebėta, kad *nors stochastinių eksperimentų konkretūs (individualūs, atskiri) rezultatai „elgiasi labai netaisyklingai“, šių eksperimentų rezultatai pakankamai ilgoje stochastinių eksperimentų serijoje yra stabilūs*.

Santykinio dažnio sąvoka įvedama 3-iaame skyrelyje, tačiau atsižvelgdami į ką tik suformuluoto statistinio stabilumo fenomeno svarbą ir norėdami ją patikslinti, užbėgsime šiek tiek į priekį.

Kartosime mūsų stochastinį eksperimentą pakankamai daug kartų ir stebėsime, kiek kartų įvyko rezultatas A . Tegu pirmuosiuose n eksperimentų A įvyko ν kartų. Tada statistinio stabilumo¹ fenomenas galėtų būti formuluojamas taip:

Didėjant stochastinių eksperimentų skaičiui n , rezultato (įvykio) A santykinis dažnis $\frac{\nu}{n}$ turi tendenciją įgyti pastovią reikšmę.

Šio statistinio stabilumo fenomeno iliustracija galėtų pasitarnauti aukščiau aprašytas monetos metymo stochastinis eksperimentas. Pažymėkime H herbo pasirodymų skaičių po pirmų n bandymų. Jau XVIII a. (kai kuriais duomenimis ir daug anksčiau) buvo pastebėta, kad santykis $\nu(H)/n$ yra stabilus ir apytiksliai lygus $\frac{1}{2}$ (žr. 1 lentelę).

¹ Šis terminas paimtas iš H. Kramero knygos [7]. R.A. Johnson ir G.K. Bhattacharyya [5] naudoja terminą „long-run stability of relative frequency“.

1 lentelė

Eksperimentatorius	Bandymų skaičius n	Herbo pasirodymų skaičius $\nu(H)$	$\nu(H)/n$
Ž. Biufonas (XVIII a.)	4040	2048	0,5080
K. Pirsonas (XIX a.)	12000	6019	0,5016
K. Pirsonas (XIX a.)	24000	12012	0,5005

Statistinio stabilumo sąvoka sąlygojo kitos fundamentalios sąvokos – *atsitiktinio įvykio tikimybės* – atsiradimą.

2. ELEMENTARIJŲ ĮVYKIŲ ERDVĖ. ATSITIKTINIS ĮVYKIS. POPULIACIJOS IR IMTIES SĄVOKOS

1. Pracitame skyrelyje išnagrinėjome stochastinio eksperimento (arba bandymo) sąvoką. Dabar tarkime, kad stochastinio eksperimento visus konkrečius rezultatus galima aprašyti aibe Ω , t.y. aibės Ω elementai pilnai charakterizuoja (pateikia pilną informaciją apie) nagrinėjamo stochastinio eksperimento stebimus rezultatus.

Stochastinio eksperimento visų galimų rezultatų aibė Ω vadinama elementariųjų įvykių erdve, o jos elementai ω_j – elementariaisiais įvykiais.

2. Išnagrinėsime kelis pavyzdžius. Stochastinio eksperimento, kurio metu vieną kartą metama moneta, elementariųjų įvykių erdvė yra $\Omega = \{H, S\}$, kur raide H pažymėtas herbo pasirodymas, o raide S – skaičiaus pasirodymas.

Jei eksperimento metu yra mėtomas lošimo kauliukas, tai

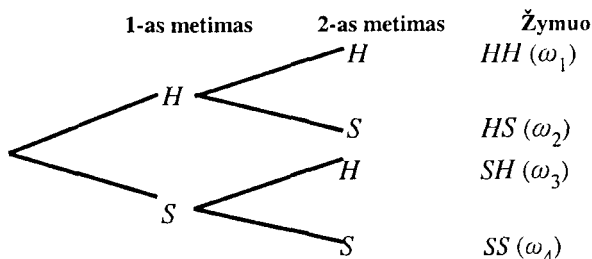
$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Atlikime kitą stochastinį eksperimentą – monetą meskime du kartus. Čia jau elementariųjų įvykių erdvė irgi yra kita:

$$\Omega = \{HH, HS, SH, SS\} = \{\omega_1, \omega_2, \omega_3, \omega_4\}.$$

Šiuo atveju patogiu elementariuosius įvykius ω_j vaizduoti *medžio diagrama* (žr. 13 psl.).

3. Išnagrinėtuose pavyzdžiuose Ω turėjo baigtinį elementų skaičių, t.y. elementariųjų įvykių aibė buvo baigtinė. Tačiau daugumoje statistikos uždavinių

*Medžio diagrama*

stochastinis eksperimentas turi be galo daug stebimų rezultatų. Išnagrinėsime porą tokių pavyzdžių.

Atliekamas toks stochastinis eksperimentas – moneta metama iki herbo pirmojo pasirodymo. Tada $\Omega = \{\omega_1, \omega_2, \dots, \omega_\infty\}$, kur

$$\omega_n = \underbrace{SS \dots S}_{n-1} H.$$

Čia elementarus įvykis ω_n reiškia, kad herbas pirmąkart pasirodys n -tojo metimo metu, o ω_∞ pažymėtas įvykis „herbas niekada nepasirodys“.

Kitas pavyzdys – Jonas ir Petras turi susitikti prie Vilniaus Katedros centrinio įėjimo laiko intervale $[T_1; T_2]$. Stochastinis eksperimentas yra toks: stebima, kada nurodytame intervale prie Katedros ateina Jonas ir kada – Petras. Tada

$$\Omega = \{(x, y) \mid T_1 \leq x \leq T_2, T_1 \leq y \leq T_2\}.$$

4. Jei aibė A yra elementariųjų įvykių erdvės Ω poaibis, tai A yra vadinamas atsitiktiniu įvykiu. Pavyzdžiui, jei moneta metama du kartus, o A – įvykis, kuris reiškia, kad bent vieną kartą pasirodė herbas, tai

$$\Omega = \{HH, HS, SS, SH\}, \quad A = \{HH, HS, SH\}.$$

Kadangi elementariųjų įvykių ω visuma Ω suteikia pilną informaciją apie stochastinio eksperimento rezultatus, tai tuo atveju, kai eksperimento konkretus rezultatas $\omega \in A$ sakome, kad įvykis A įvyko, o kai $\omega \notin A$ – neįvyko.

Aibės A elementai ω vadinami elementariaisiais įvykiais, palankiais įvykiui A .

5. Taigi, įvykio A atžvilgiu elementariųjų įvykių erdvę Ω galima suskaidyti į dvi dalis – į aibę A ir jos papildinį $\bar{A}$.

Pati aibė Ω (t.y. elementariųjų įvykių erdvė), jei yra nagrinėjama kaip atsitiktinis įvykis, charakterizuojama tuo, kad, nepriklausomai nuo stochastinio eksperimento rezultato, šis įvykis Ω būtinai įvyksta.

Šio teiginio įrodymas labai paprastas, nes pagal elementariųjų įvykių erdvės apibrėžimą jokie kiti stochastinio eksperimento rezultatai, išskyrus tuos, kurie aprašomi elementais $\omega \in \Omega$, yra negalimi. Štai kodėl aibė Ω yra vadinama būtinuoju įvykiu¹.

6. Simboliu $\emptyset$ priimta žymėti aibę, neturinčią nei vieno elemento. Ji vadinama tuščiąja aibe. Taigi, tuščioji aibė $\emptyset$ yra aibės Ω „priešingybė“.

Jei aibę $\emptyset$ nagrinėti kaip atsitiktinį įvykį, tai toks įvykis stochastiniame eksperimente niekada neįvyksta, todėl vadinamas negalimuoju įvykiu².

7. Skyrelio pabaigoje pateiksime keletą terminologinių pastabų.

Pagal Lietuvos standartą [8], **populiacija**³ – tai visų stochastiniame eksperimente nagrinėjamų elementų aibė.

Ėmimo vienetas – tai populiacijos elementas.

Pagal tą patį standartą, **imtis**⁴ (arba **atranka**) – tai vienas ar daugiau ėmimo vienetų, paimtų iš populiacijos, siekiant gauti informaciją apie populiaciją. **Imties dydis**⁵ – tai ėmimo vienetų skaičius imtyje.

Terminą *atranka* dažnai naudoja ekonomistai. Svarbu pabrėžti, kad tai atsitiktinio ėmimo sinonimas. Būtų neteisinga manyti, kad tai specialus pasirinkimas, kai pagal kažkokį požymį kažkas yra atrenkama.

Šiuolaikiniai statistikos vadovėliai pateikia dar vieną aibės Ω apibrėžimą – *imčių erdvė*⁶ (žr., pavyzdžiui, R.A. Johnson, G.K. Bhattacharyya [5]). Pateiksime šį ir giminingus apibrėžimus pilnai.

Imčių erdvė, susijusia su stochastiniu eksperimentu, vadinama šio eksperimento visų galimų skirtingų rezultatų (baigčių) aibė.

Kiekvienas šio eksperimento rezultatas vadinamas **elementariąja baigtimi**, arba **elementariuoju (paprastuoju) įvykiu**, arba **imčių erdvės elementu**, arba **stebiniu**.

¹ Angl. *certain event*, rus. *достоверное событие*.

² Angl. *impossible event*, rus. *невозможное событие*.

³ Angl. *certain event*, rus. *генеральная совокупность*.

⁴ Angl. *sample*, rus. *выборка*.

⁵ Angl. *sample size*, rus. *объем выборки*.

⁶ Angl. *sample space*, rus. *выборочное пространство*.

Atsitiktinių įvykiu vadinama elementariųjų baigčių, turinčių apibrėžtas savybes, aibė.

3. STATISTINIŲ DUOMENŲ APDOROJIMAS. SANTYKINIS DAŽNIS. DAŽNIŲ SKIRSTINYS

1. Statistinius duomenis galima suskirstyti į du pagrindinius tipus:

(A) *kokybiniai duomenys*¹ (arba *duomenys pagal kategorijas* (klases, grupes)),

(B) *kiekybiniai duomenys*² (arba *skaitiniai duomenys*).

2. Iš pradžių duomenys paprastai apdorojami kokybiškai, t.y. suskirstomi į *kategorijas* (klases, grupes) pagal požymius. Pavyzdžiui, nustatomos amžiaus grupės (iki 19 metų, 20–24, 25–29, 30–39, 40–49, 50 ir daugiau *ar pan.*), plaukų spalva (blondinas, brunetas *ir pan.*), užimtumas (bedarbis, užimtas darbu), mirtingumo priežastys (infekcinės ir parazitinės ligos, kraujo apytakos sistemos ligos, piktybiniai augliai *ir pan.*).

Štai paprasčiausias kokybinių duomenų tipo pavyzdys.

Pavyzdys. Patikrinus 20 magistrantų kraujo grupę (I, II, III arba IV), buvo gauti tokie duomenys:

I	II	I	I	II	IV	I	III	I	I
III	I	II	II	III	I	II	II	I	II

3. Išreikškus rezultatus kiekybinėmis (t.y. skaitinėmis) charakteristikomis pagal kategorijas, duomenys tuo pačiu apdorojami kiekybiškai. Kiekvienas *stebinys* turi būti priskirtas tik vienai iš kelių kategorijų. Tokie duomenys paprastai pateikiami *dažnių lentelės* pavidalu, kuri parodo kiekvienos kategorijos elementų skaičių (*dažnį*). Šio dažnio santykis su visų stebinių skaičiumi vadinamas *santykiniu dažniu*:

$$\text{Kategorijos santykinis dažnis} = \frac{\text{Kategorijos dažnis}}{\text{Visų stebinių skaičius}}$$

Dažnių lentelę pavyzdyje su magistrantų kraujo grupėmis galima sudaryti taip, kaip tai atlikta 2 lentelėje.

4. Dabar pacituosime Lietuvos standartą.

¹ Angl. *qualitative data*, rus. *качественные данные*.

² Angl. *quantitative data*, rus. *количественные данные*.

2 lentelė

Kraujo grupė	Dažnis	Santykinis dažnis
I	9	$9/20 = 0,45$
II	7	$7/20 = 0,35$
III	3	$3/20 = 0,15$
IV	1	$1/20 = 0,05$
Viso	20	1,00

Požymis¹ – tai savybė, padedanti identifikuoti, atskirti ar klasifikuoti populiacijos individus.

Stebinys² – tai požymio reikšmė, gauta kaip atskiros stebėjimo rezultatas.

Klasė, grupė: (1) kokybinio požymio atveju – grupė elementų, turinčių tam tikras bendras savybes. Grupės nesikerta ir apima visą populiaciją;

(2) kiekybinio požymio atveju – kiekvienas iš nesikertančių intervalų, į kuriuos padalytas visas kitimo intervalas.

Dažnis³ – tai tam tikro tipo įvykių skaičius arba į tam tikrą klasę patekusių stebinių skaičius.

Santykinis dažnis⁴ – tai dažnis, padalytas iš viso bandymų ar stebinių skaičiaus.

Dažnių skirstinys – tai požymio reikšmių ir jų dažnių sąsaja. (Skirstinys gali būti grafiškai pateiktas kaip histograma, stulpelinė diagrama, ... arba kaip dažnių lentelė).

Pavyzdys. Susituokusių 1997 metais Lietuvos vyrų dažniai pagal amžių pavaizduoti 3 lentelėje.

5. Aptarsime stulpelinės diagramos ir histogramos sąvokas, kurios buvo minimos iliustruojant dažnių skirstinio sąvoką.

Stulpelinė diagrama⁵ – tai kiekybinio požymio dažnių grafinė išraiška, susidedanti iš vienodo pločio stulpelių, kurių aukščiai proporcingi dažniams.

Atkreipsime dėmesį į svarbią detalę – stulpelių pločiai yra vienodi, nors nagrinėjamos klasės kiekybine prasme gali būti skirtingos. Mūsų pavyzdyje vyrų klasėje nuo 20 iki 24 metų amžiaus skirtumas gali siekti 5 metus, klasėje

¹ Angl. characteristic, rus. признак.

² Angl. observed value, rus. наблюдение.

³ Angl. frequency, rus. частота.

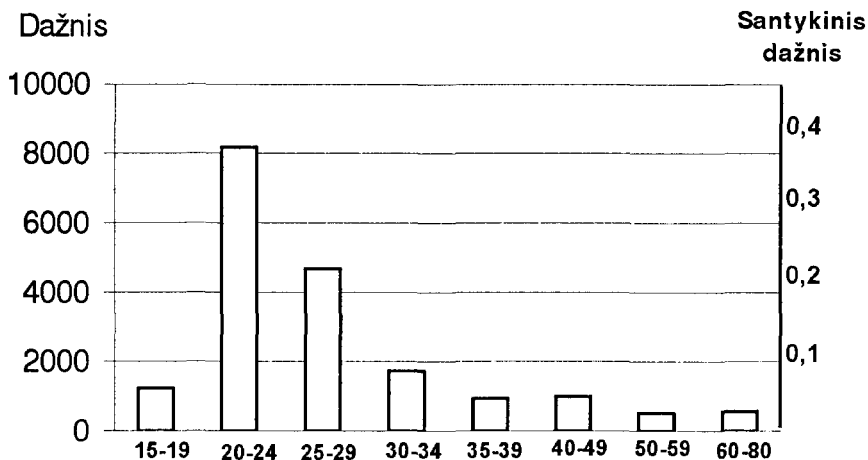
⁴ Angl. relative frequency, rus. относительная частота.

⁵ Angl. bar chart, bar diagram.

3 lentelė

Kategorija (anižiaus klasė, grupė)	Dažnis (vyrų skaičius amžiaus klasėje)	Santykinis dažnis
15–19	1225	$1225/18796 = 0,0652$
20–24	8181	$8181/18796 = 0,4353$
25–29	4672	$4672/18796 = 0,2486$
30–34	1729	$1729/18796 = 0,0920$
35–39	928	$928/18796 = 0,0494$
40–49	995	$995/18796 = 0,0529$
50–59	512	$512/18796 = 0,0272$
60–80	554	$554/18796 = 0,0294$
Iš viso	18796	1,0000

nuo 40 iki 49 metų šis skirtumas gali siekti 10 metų, o klasėje nuo 60 iki 80 – net 21 metus. Taigi, klasių plotis (ši sąvoka aiškinama žemiau) gali būti skirtingas, tačiau diagramos stulpelių plotis vienodas. Tai naudinga tuo atveju, kai pagrindinis dėmesys koncentruojamas į dažnius. Mūsų pavyzdyje apie susituokusius vyrus stulpelinė diagrama yra tokia:



1 pav. Susituokusių 1997 m. Lietuvos vyrų stulpelinė diagrama pagal amžių

6. Grįžkime prie paskutinio pavyzdžio dažnių lentelės. Blokai 20–24, 25–29, 30–34 ir pan. yra vadinami *klasių intervalais*. Apatinis klasės intervalo

galas vadinamas *klasės apatine riba*, o viršutinis galas – *klasės viršutine riba*. Šiose ribose turi tilpti nagrinėjamos klasės intervalas. Taigi, mūsų pavyzdyje 15, 20, 25, 30, 35, 40, 50, 60 yra klasės apatinės ribos, o $19 + 1$, $24 + 1$, $29 + 1$, $34 + 1$, $39 + 1$, $49 + 1$, $59 + 1$, $80 + 1$ – viršutinės ribos. Kodėl plius 1? Jei žmogui yra 24 metai ir 364 dienos, t.y. beveik 25 metai, mes jį vis tiek priskiriame klasei 20–24. Tai daroma tam, kad vienas ir tas pats žmogus būtų priskirtas tik vienai klasei.

Tokioms situacijoms Lietuvos standarte yra skirtos dvi pastabos: 1) turi būti patikslinta, kurios iš šių dviejų ribų (apatinė ar viršutinė) priskiriamos klasei; 2) jei įmanoma, klasės ribos neturi sutapti su galima reikšme.

Mūsų pavyzdyje yra patikslinta, kad apatinė riba priklauso klasei, o viršutinė – ne. Be to, klasės viršutinė riba nėra jos galima reikšmė. Taigi, nėra dviprasmybės ir painiavos.

Klasės plotis¹ – tai *klasės viršutinės ribos ir apatinės ribos skirtumas*. Nagrinėjamajame pavyzdyje klasių 15–19, 20–24, 25–29, 30–34, 35–39 pločiai yra lygūs 5, klasių 40–49 ir 50–59 pločiai – 10, o klasės 60–80 plotis yra 21.

7. Dabar turime visas reikalingas sąvokas histogramos apibrėžimui.

Histograma – tai *kiekybinio požymio dažnių skirstinio grafinė išraiška, susidedanti iš keletu besiribojančių stačiakampių, kurių kiekvieno pagrindas lygus klasės pločiui, o plotas lygus klasės santykiniam dažniui*.

Kadangi visų stačiakampių plotas yra lygus klasių santykinių dažnių sumai, tai gauname tokią taisyklę: **histogramos plotas lygus 1**.

Lietuvos standarte pateikiamas histogramos apibrėžimas skiriasi tik paskutiniaisiais žodžiais: „... o plotas proporcingas klasės dažniui“. Taip apibrėžus histogramą, jos plotas nebūtinai lygus 1, nors sąvokos esmė išlieka ta pati. Statistikos vadovėliuose yra patogesnis pirmasis apibrėžimas, nes asocijuojasi su tankio sąvokos įvedimu.

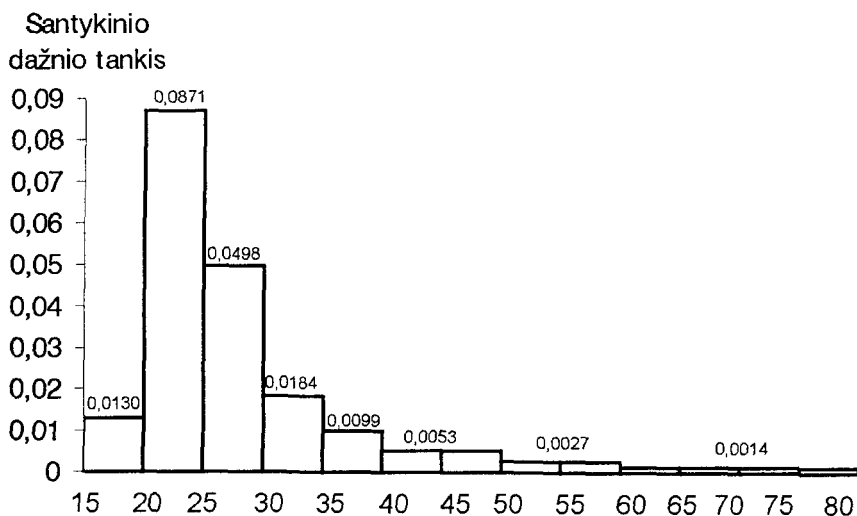
Mūsų pavyzdyje apie susituokusius vyrus histograma yra pavaizduota 2 pav.

Skaičius virš stačiakampio reiškia jo aukštį, kuris apskaičiuojamas pagal tokią formulę:

$$\text{Aukštis} = \frac{\text{Klasės santykinis dažnis}}{\text{Klasės plotis}}$$

Pastaroji trupmena dar yra vadinama *klasės santykinio dažnio tankiu*.

¹ Angl. *class width, class size*.



2 pav. Susituokusių 1997 m. Lietuvos vyrų histograma pagal amžių

8. Kadangi pasirinktame pavyzdyje kalbama vien apie vedusius vyrus, būtų neteisinga nutyliėti informaciją apie ištekęsias 1997 metais Lietuvos moteris. Ši informacija pateikiama 4 lentelėje.

4 lentelė

Amžiaus klasė	Ištekęsųjų moterų skaičius amžiaus klaseje
15–19	3973
20–24	8128
25–29	3044
30–34	1252
35–39	799
40–49	789
50–59	460
60–80	353
Iš viso	18796

Siūlome savarankiškai sudaryti dažnių lentelę, stulpelinę diagramą ir histogramą.

4. TIKIMYBINIO MATO ĮVEDIMAS. ATSITIKTINIO ĮVYKIO TIKIMYBĖ

1. Išnagrinėję statistinio stabilumo fenomeną ir turėdami atsiktinio įvykio bei dažnio sąvokas, galime kalbėti apie tikimybinio mato įvedimą.

Atsitiktiniu įvykiu pavadiname elementariųjų baigčių (stebinių), turinčių apibrėžtas savybes, aibę. *Dažnis* – tai tam tikro tipo atsitiktinių įvykių skaičius stochastinių eksperimentų serijoje.

Tarkime, kad kalbama apie kažkokį įvykį A , kurio dažnis n stochastinių eksperimentų serijoje yra $n(A)$. 1-ajame skyrelyje išnagrinėtą *statistinio stabilumo fenomeną* dabar galime konkretizuoti.

Atsitiktinio įvykio A santykinis dažnis $n(A)/n$ n stochastinių eksperimentų serijoje turi tendenciją, esant pakankamai dideliui n , artėti prie pastovios reikšmės, kurią pažymėsime $P(A)$. Skaičius $P(A)$ vadinamas atsitiktinio įvykio A tikimybe.

2. Atsitiktinio įvykio ir jo tikimybės sąvokas mes kol kas nagrinėjame intuityviame lygmenyje. Griežtus matematinius apibrėžimus ir teiginius, pradedant aksiomomis, nagrinėsime II-ojoje vadovėlio dalyje „Matematiniai statistikos metodai“.

3. Priskirdami kiekvienam atsitiktiniam įvykiui neneigiamą skaičių (santykis $n(A)/n$ visada neneigiamas, taigi, ir šio santykio riba, kai $n \rightarrow \infty$, yra neneigiama), mes turime teisę tvirtinti, kad šį įvykį mes „išmatuojame“.

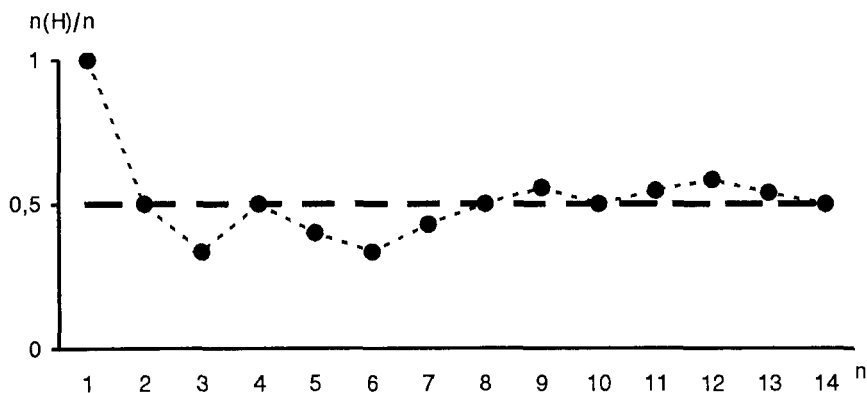
Analogiška situacija susidaro tuo atveju, kai parduotuvėje tenka išreikšti pinigine išraiška (t.y. „išmatuoti litais“), tarkime, bananų kekės vertę. Kaip mes tai darome? Sveriamė bananus elektroninėmis svarstyklėmis, kuriose taip pat nustatoma 1 kg bananų kaina litais. Taigi, kiekvienam bananui, kiekvienai kekei priskiriamas teigiamas skaičius, reiškiantis jo arba jos vertę litais.

Tokiu būdu galima tvirtinti, kad kiekviena bananų kekė yra „išmatuojama“, jai yra priskiriamas skaičius, sutampantis su jos verte litais. Tačiau reikia turėti omenyje, kad sąvokos *matas* ir *tikimybinis matas* skiriasi tuo, kad pastarasis yra *normuotas matas*, jo reikšmė negali viršyti vienetą.

4. Santykinio dažnio $n(A)/n$ artėjimą prie pastovios reikšmės pailiustruosime paprasčiausiu pavyzdžiu, kurį jau nagrinėjome 1-ajame skyrelyje. Tai monetos mėtymo stochastinis eksperimentas. Paimkime monetą ir atlikime keletą metimų. Šio vadovėlio autorius, atlikęs monetos mėtymo eksperimentą 14 kartų, gavo rezultatus, kurie pavaizduoti 5 lentelėje.

5 lentelė

Eksperimentų skaičius n	Herbų pasirodymų skaičius $n(H)$ (dažnis $n(H)$)	Santykinis dažnis $n(H)/n$
1	1	1
2	1	1/2
3	1	1/3
4	2	1/2
5	2	2/5
6	2	1/3
7	3	3/7
8	4	1/2
9	5	5/9
10	5	1/2
11	6	6/11
12	7	7/12
13	7	7/13
14	7	1/2



3 pav. Santykinio dažnio $n(H)/n$ svyravimų amplitudžių mažėjimas bei statistinio stabilumo fenomenas monetos mėtymo stochastiniame eksperimente

Grafiškai šio stochastinio eksperimento rezultatus galima pavaizduoti 3 pav.

Iš grafiko matome, kad, augant n , svyravimų amplitudė mažėja¹, o san-

¹ Prisiminkime taip pat Biufono–Pirsono stochastinius eksperimentus, išnagrinėtus 1-ajame skyrelyje.

tykinis dažnis $n(H)/n$ artėja prie $1/2$.

5. Dabar galime suformuluoti tokią svarbią išvadą.

Kiekvieną kartą, kai tvirtinama, kad atsitiktinio įvykio A tikimybė atliekame stochastiniame eksperimente lygi $P(A)$, šio teiginio prasmė yra tokia: praktiškai neabejotina, kad atsitiktinio įvykio A santykinis dažnis ilgoje stochastinio eksperimento kartojimų sekoje apytiksliai lygus tikimybei $P(A)$.

6. Nesunku įsitikinti, kad bet kokio atsitiktinio įvykio A tikimybė $P(A)$ tenkina šias savybes:

$$1) 0 \leq P(A) \leq 1; \quad 2) P(A) = \sum_{\omega \in A} P(\omega); \quad 3) P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1.$$

5. ATSITIKTINIAI DYDŽIAI. TIKIMYBĖS MASĖS FUNKCIJA. TIKIMYBINĖ HISTOGRAMA

1. Automobilių skaičius, kuriuos per dieną aptarnauja autoservisas, prarasto svorio kiekis po dietos kurso pabaigimo, laisvų vietų skaičius reisiniame keleiviniame lėktuve ir t.t. – visa tai yra įvairių stochastinių eksperimentų rezultatų (baigčių) skaitinės išraiškos. Sukoncentravę savo dėmesį į šių baigčių skaitines reikšmes, įvesime *atsitiktinio dydžio* sąvoką.

Atsitiktinis dydis yra funkcija, suteikianti kiekvienai stochastinio eksperimento baigčiai skaitinę reikšmę¹.

2. Išnagrinėsime pavyzdį. Tarkime, kad į erdvę paleisti trys palydovai. Raide X pažymėkime palydovų, pasiekusių orbitą (t.y. savo tikslą), skaičių. Aprašysime atsitiktinį dydį X .

Tuo tikslu pirmiausia aprašysime visas stochastinio eksperimento baigtis (imčių erdvę), o po to nustatysime atsitiktinio dydžio X skaitines reikšmes šios erdvės taškuose. Raide s pažymėkime baigtį, kai palydovas pasiekė orbitą sėkmingai, o raide n – baigtį, kai palydovas nepasiekė orbitos. Tada atsitiktinį dydį X – funkciją iš baigčių erdvės (imčių arba elementariųjų įvykių erdvės) į natūrinių (įskaitant nulį) skaičių aibę – galima aprašyti tokia lentele:

Stochastinio eksperimento baigtis ω_j	sss	ssn	sns	nss	nns	nsn	snn	nnn
Atsitiktinio dydžio X reikšmė $x_j = X(\omega_j)$	3	2	2	2	1	1	1	0

¹ Antroje vadovėlio dalyje šis apibrėžimas bus patikslintas.

Taigi, atsitiktinis dydis X įgyja keturias skirtingas reikšmes: 0, 1, 2 ir 3.

3. Pagal Lietuvos standartą, *atsitiktinis dydis, kuris gali įgyti tik izoliuotas reikšmes, vadinamas diskrečiuoju*. Ką tik išnagrinėtas pavyzdys su palydovais – diskrečiojo atsitiktinio dydžio pavyzdys.

Atsitiktinis dydis, kuris gali įgyti kiekvieną reikšmę iš baigtinio ar begalinio intervalo, vadinamas tolydžiuoju. Žinoma, bet kurie matavimai yra atliekami su tam tikru tikslumu, todėl praktikoje tolydi reikšmių skalė yra tam tikra abstrakcija. Žmogaus svoris, maksimali dienos temperatūra, elektros lemputės darbo laiko trukmė – visa tai yra tolydžių atsitiktinių dydžių pavyzdžiai.

4. Atsitiktinio dydžio X reikšmių lentelė suteikia galimybę žinoti visas stochastinio eksperimento baigtis ir atsitiktinio dydžio X įgyjamas skaitines reikšmes. Prisiminę tikimybės sąvoką, galime nustatyti skirtingų atsitiktinio dydžio X įgyjamų reikšmių pasirodymo šansus. Šiuo tikslu yra įvedama atsitiktinio dydžio X *tikimybės masės funkcija*, kurios apibrėžimą pateiksime pagal Lietuvos standartą.

Tikimybės masės funkcija¹ – tai funkcija, priskirianti kiekvienai diskrečiojo atsitiktinio dydžio X reikšmei x_i tikimybę $f(x_i)$, kad X įgis reikšmę x_i :

$$f(x_i) = P(X = x_i).$$

Kartais, norėdami supaprastinti žymėjimą, riestinius skliaustus praleisime: $f(x_i) = P(\{X = x_i\}) = P(X = x_i)$. Žinoma, tiksliausia būtų rašyti $f(x_i) = P(\{X = x_i\}) = P(\{\omega \mid X(\omega) = x_i\})$.

5. Pratešime antrame punkte pateikto pavyzdžio apie atsitiktinio dydžio X , reiškiančio palydovų, sėkmingai pasiekusių orbitą, nagrinėjimą. Lentelės pavidalu pateiksime šio atsitiktinio dydžio tikimybės masės funkciją.

Mes matėme, kad šis statistinis eksperimentas turi aštuonias baigtis, t.y. šio eksperimentų elementariųjų įvykių erdvė sudaryta iš 8 elementų: $\omega_1 = sss$, $\omega_2 = ssn$, $\omega_3 = sns$, $\omega_4 = nss$, $\omega_5 = nns$, $\omega_6 = nsn$, $\omega_7 = snn$, $\omega_8 = nnn$. Tarkime, kad yra išbandoma principingai nauja kosminių palydovų karta, todėl bet kuri stochastinio eksperimento baigtis yra vienodai galima, taigi, jos tikimybė yra $1/8$.

¹ Angl. *probability mass function*, sutrumpintai – *probability function*, dažnai – *probability distribution*.

Atsitiktinis įvykis $\{X = 0\}$ gali būti realizuojamas vieninteliame elementariųjų įvykių erdvės taške ω_8 , t.y. kai stochastinio eksperimento baigtis – visiška nesėkmė (visi trys palydovai nepasiekia orbitos). Taigi,

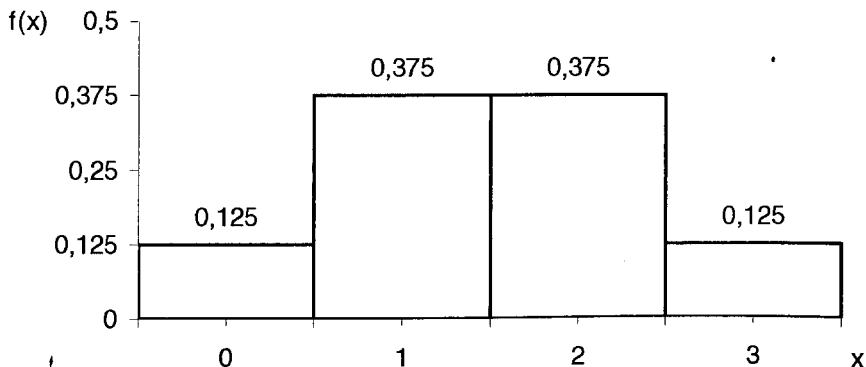
$$P(X = 0) = P\{\omega \mid X(\omega) = 0\} = P(\{\omega_8\}) = 1/8.$$

Analogiškai, atsitiktiniai įvykiai $\{X = 1\}$, $\{X = 2\}$ ir $\{X = 3\}$ turi atitinkamai 3, 3 ir 1 elementą bei tikimybės $3/8$, $3/8$ ir $1/8$. Gauname tokias diskretaus dydžio X tikimybės masės funkcijos $f(x)$ reikšmes, kurios pavaizduotos 6 lentelėje.

6 lentelė

Atsitiktinio dydžio X reikšmė x_i	Tikimybė $f(x_i)$
$x_1 = 0$	$f(x_1) = 1/8$
$x_2 = 1$	$f(x_2) = 3/8$
$x_3 = 2$	$f(x_3) = 3/8$
$x_4 = 3$	$f(x_4) = 1/8$
Iš viso	1

6. Grafiškai tikimybės masės funkciją yra patogu vaizduoti histograma (žr. 3-ąjį skyrelį), kuri vadinama *tikimybine histograma*. Joje stačiakampių pagrindų centrais pasirenkami taškai x_i , o kiekvieno tokio stačiakampio plotas lygus $f(x_i)$:



4 pav. Tikimybės masės funkcijos histogramos (tikimybinių histogramos) pavyzdys

7. Reziumuodami skyrelyje išnagrinėtą medžiagą, suformuluosime **tikimybės masės funkcijos** $f(x)$ savybes:

- 1) *Tikimybė, kad atsitiktinis dydis įgis reikšmę x_i , visada yra tarp 0 ir 1, t.y.*

$$0 \leq f(x_i) \leq 1;$$

- 2) *Diskretaus atsitiktinio dydžio, įgyjančio reikšmes $x_1, \dots, x_n$, tikimybės masės funkcijos reikšmių suma lygi 1, t.y.*

$$\sum_{i=1}^n f(x_i) = 1.$$

6. POPULIACIJOS VIDURKIS. DIDŽIŲJŲ SKAIČIŲ DĖSNIS. DRAUDIMO IR LOŠIMO MODELIAI

1. Pradėkime nuo paprasčiausio statistinio duomenų rinkimo. Tuo tikslu 20 kartų meskime lošimo kauliuką. Šių eilučių autorius gavo tokius duomenis:

3	6	3	5	4	1	2	4	4	2
3	4	6	5	1	3	6	6	3	1

Šių stebinių *aritmetinis vidurkis* $\bar{x}$, dar vadinamas *imties vidurkiu*, apskaičiuojamas taip:

$$\bar{x} = \frac{\text{stebinių reikšmių suma}}{\text{stebinių skaičius}} = \frac{72}{20} = 3,6.$$

2. Tačiau šį vidurkį mes galėjome apskaičiuoti kitaip. O būtent, galėjome apskaičiuoti kiekvienos kubiuko pusės pasirodymo dažnį ir, apskaičiavę santykinį dažnį, vidurkį gauname taip:

$$\bar{x} = 1 \cdot \left(\frac{3}{20}\right) + 2 \cdot \left(\frac{2}{20}\right) + 3 \cdot \left(\frac{5}{20}\right) + 4 \cdot \left(\frac{4}{20}\right) + 5 \cdot \left(\frac{2}{20}\right) + 6 \cdot \left(\frac{4}{20}\right) = 3,6.$$

Šį skaičiavimo būdą galima iliustruoti tokia formule:

$$\text{Imties vidurkis } \bar{x} = \sum (\text{reikšmė} \times \text{santykinis dažnis}).$$

3. Įsivaiduokime, kad kubiuką mes mėtome ne 20, o labai daug kartų. Tada santykinis dažnis, kaip matėme 4-ajame skyrelyje, artės prie vienos kubiuko sienelės pasirodymo tikimybės, t.y. prie $1/6$. Vidurkis tokiu atveju apskaičiuojamas taip:

$$1 \cdot \left(\frac{1}{6}\right) + 2 \cdot \left(\frac{1}{6}\right) + \dots + 6 \cdot \left(\frac{1}{6}\right) = \sum (\text{reikšmė} \times \text{tikimybė}) = 3,5.$$

Šio pavyzdžio ir statistinio stabilumo fenomeno (kurį išnagrinėjome 1-ajame skyrelyje) pagrindu natūralu *atsitiktinio dydžio X vidurkį* apibrėžti taip:

$$\sum (\text{reikšmė} \times \text{tikimybė}) \text{ arba } \sum x_i f(x_i),$$

kur x_i pažymėtos skirtingos atsitiktinio dydžio X reikšmės, o $f(x)$ – tikimybės masės funkcija.

4. Atsitiktinio dydžio X vidurkis dar yra vadinamas **populiacijos vidurkiu** arba **matematinė viltimi** ir žymimas μ arba EX .

O štai Lietuvos standarto apibrėžimas:

Diskrečiojo atsitiktinio dydžio X , įgyjančio reikšmes x_i su tikimybėmis p_i , vidurkis (kai jis egzistuoja) yra

$$\mu = EX = \sum x_i p_i;$$

čia sumuojama pagal visas X įgyjamas reikšmes x_i .

5. Apibendrinkime kubiuko mėtymo eksperimentą, pasirinkdami bet kokią stochastinį eksperimentą. Kartokime jį daug kartų, kaskart skaičiuodami jo stebinių aritmetinį vidurkį $\bar{x}$. Galima įsitikinti, kad ši apskaičiuotoji reikšmė $\bar{x}$ po daugelio eksperimento kartojimų yra artima teoriniam vidurkiui μ . Tai taip vadinamas *didžiųjų skaičių dėsnis*¹. Suformuluosime jį atskirai.

Didžiųjų skaičių dėsnis. *Praktiškai neabejotina, kad didelio skaičiaus nepriklausomų populiacijos stebinių aritmetiniai vidurkiai artėja prie šios populiacijos vidurkio μ .*

Norėdami iliustruoti šio dėsnio veikimą, atlikime tokį eksperimentą². Modeliuokime 400 atsitiktinio dydžio su nuliniu vidurkiu populiacijos³ stebinių.

¹ Angl. *law of large numbers*, rus. *закон больших чисел*.

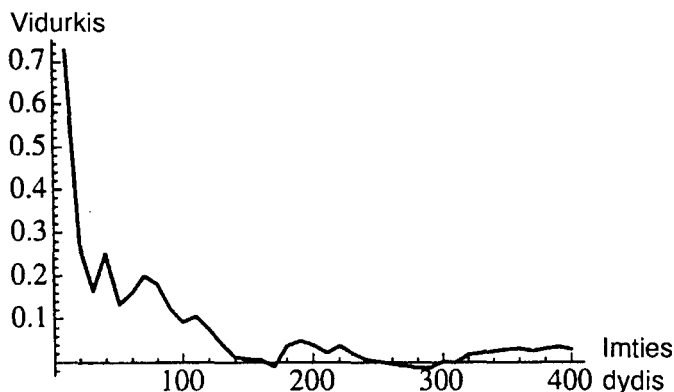
² Eksperimento autorius – K.J. Hastings [3].

³ K.J. Hastings nagrinėjo standartinio normaliojo atsitiktinio dydžio populiaciją.

Kiekvieną kartą, gavę 10 naujų stebinių, iš naujo apskaičiuojame visos imties vidurkį. Tai reiškia, kad skaičiuojamos šios reikšmės:

$$\bar{x}_{10} = (x_1 + \dots + x_{10})/10, \quad \bar{x}_{20} = (x_1 + \dots + x_{20})/20, \dots, \\ \bar{x}_{400} = (x_1 + \dots + x_{400})/400.$$

Gauname tokį grafiką:



5 pav. Standartinio normaliojo atsitiktinio dydžio populiacijos stebinių modeliavimo rezultatai, iliustruojantys didžiųjų skaičių dėsnio veikimą

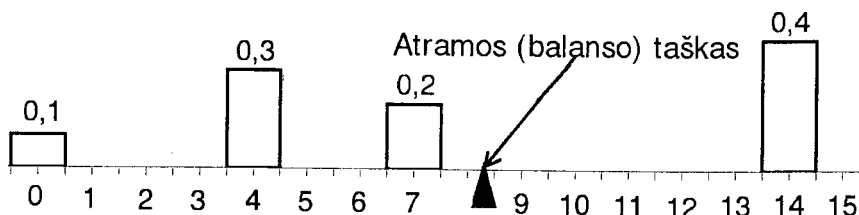
Išvada aiški: didėjant stebinių skaičiui, jų aritmetinis vidurkis artėja į nulį, t.y. į populiacijos vidurkį $\mu = 0$.

6. Fizikinė vidurkio interpretacija. Vidurkį galima interpretuoti kaip atsitiktinio dydžio „svorio centrą“. Įsivaiduokime svorius, kurių skaitinė išraiška kilogramais lygi diskrečiojo atsitiktinio dydžio įgyjamų reikšmių tikimybėms. Išdėliokime juos ant svertinių sūpuoklių atstumais, proporcingais įgyjamoms reikšmėms. Norint, kad tokios sūpuoklės įgytų pusiausvyrą, atramos tašką reikia lokalizuoti tiksliai toje vietoje, kurios koordinatė lygi atsitiktinio dydžio vidurkiui.

Pailiustruosime tai pavyzdžiu. Tegu atsitiktinis dydis X įgyja reikšmes 0, 4, 7 ir 14 su tikimybėmis 0,1; 0,3; 0,2 ir 0,4. Tada

$$\mu = EX = 0 \cdot 0,1 + 4 \cdot 0,3 + 7 \cdot 0,2 + 14 \cdot 0,4 = 8,2$$

o svertines sūpuokles su svoriais 0,2; 0,3; 0,2 ir 0,4 grafiškai galima pavaizduoti taip:



6 pav. Fizikinė vidurkio interpretacija – svertinės sąpuoklės

7. Draudimo modelis. Daugelis tikimybinių sąvokų, įskaitant vidurkio arba matematinės vilties idėją, kilo studijuojant azartinius lošimus. Palaipsniui šios sąvokos ir idėjos buvo taikomos tose srityse, kuriose ženklia įtaką turėjo rizikos faktoriai. Pirmiausia, vystantis prekybai, jos rado pritaikymą paprasčiausiuose draudimo modeliuose. Vieno tokio modelio pavyzdį išnagrinėsime dabar, o sekančiame punkte išnagrinėsime azartinio lošimo modelio pavyzdį.

Draudimo kompanija išmoka klientui 5 000 Lt sumą, jei kruizo metu jis apvagiamas arba patiria kitus nuostolius. Jei šių nuostolių rizika įvertinama santykiu 1:250, tai kokia turi būti kliento draudimo įmoka?

Pirmiausia pastebėsime, jog tikimybė, kad kompanijai teks išmokėti 5 000 Lt sumą, yra $1/250=0,004$. Tegu X yra atsitiktinis dydis, reiškiantis kompanijos išmokos x klientui sumą. Tada šio dydžio tikimybės masės funkcija $f(x)$ yra tokia:

Išmoka x_i	Tikimybė $f(x_i)$
5000 Lt	0,004
0 Lt	0,996

Apskaičiuojame vidurkį:

$$EX = 5000 \times 0,004 + 0 \times 0,996 = 20 \text{ Lt}.$$

Taigi, vidutinė suma, kurią kompanija išmoka klientui, yra 20 Lt. O tai reiškia, kad būtų pagrįsta iš kliento prašyti 20 Lt didumo draudimo įmoką.

Tik tokiu atveju, atsižvelgiant į statistinio stabilumo fenomeną, ilgoje draudimų serijoje kompanija neturės nei nuostolio, nei pelno.

Tai teorinis modelis. Čia neatsižvelgta į jokiais kitas draudimo kompanijos išlaidas. Praktikoje draudimo įmoka turi būti didesnė, atsižvelgiant į administravimo išlaidas ir planuojamą pelną.

8. **Ar galima išlošti, žaidžiant kazino¹?** „Keistas klausimas. Žinoma, taip!“ – atsakys besišypsantis lošėjas, išlošęs stambia pinigų sumą. Bet kitokios mintys kyla, žvelgiant į prasilošusių žaidėjų susirūpinusius veidus. Kokia gi yra šio žaidimo matematinė viltis?

Ruletės ratas turi 18 raudonų, 18 juodų ir 2 žalius sektorius. Pastatykime 10 dolerių sumą, pavyzdžiui, į raudoną sektorių. Tegu X yra atsitiktinis dydis, reiškiantis išloštą pinigų sumą. Tada

$$EX = 10 \cdot \frac{18}{38} + (-10) \cdot \frac{20}{38} = -0,526.$$

Neigiamas matematinės vilties EX ženklas reiškia, kad kiekvienai pastatytai 10 dolerių sumai mes turime tikėtis prarasti 52,6 cento. Didelį skaičių kartų kartojant lošimą, laimėjimų santykinis dažnis artėja prie tikimybės $18/38$, o pralaimėjimų santykinis dažnis artėja prie tikimybės $20/38$. Tai reiškia, kad žaidėjas nemažą pinigų sumą paliks kazino. Ir tai suprantama, nes kaip galėtų klestėti kazino biznis priešingu atveju?

Beje, Lietuvoje šis biznis nėra legalizuotas, todėl ir statomos sumos pateikiamos doleriais, o ne litais. Mat potencialus lošėjas turėtų vykti lošti į Latviją ar kitą užsienio šalį ir neužmiršti, kad bet kuriuose lošimo namuose išlošiamų pajamų matematinė viltis yra neigiama.

7. DISPERSIJA KAIP SKLAIDOS MATAS. ČEBYŠEVO TAISYKLĖ DUOMENŲ ANALIZEI

1. Tarkime, kad $x_1, x_2, \dots, x_n$ yra stochastinio eksperimento stebinių reikšmės. Šios *imties vidurkiu*² praeitame skyrelyje pavadiname jos reikšmių aritmetinį vidurkį $\bar{x}$,

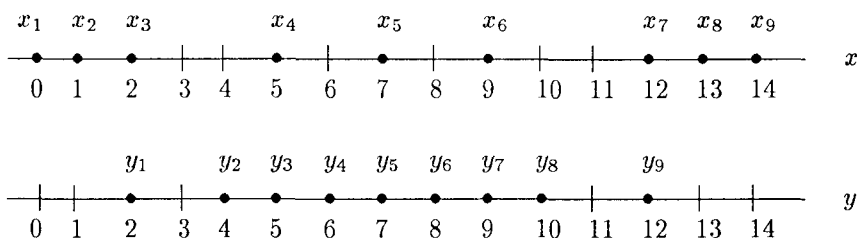
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

¹ Dėstoma pagal R.A. Johnson, G.K.Bhattacharyya [5] autorinį pavyzdį.

² Angl. *sample mean*, rus. *выборочное среднее*.

„Matematikos terminų žodyne“ [10] $\bar{x}$ vadinamas *empiriniu vidurkiu*. Matome, kad tai centro matas, kurio fizikinė interpretacija – svorio centras.

2. Tačiau be duomenų centro nustatymo, bet kuri duomenų aprašomoji analizė privalo turėti kitimo (sklaidos) apie šį centrą skaitinę išraišką, t.y. *kitimo (sklaidos) matą*. Dvi stebinių aibės gali turėti tas pačias centro pozicijas, bet skirtingą sklaidos apie šį centrą pavidalą. Pavyzdžiui, taškai viršutiniame paveiksle yra daugiau išsklaidyti, negu taškai apatiniame paveiksle, nors jų centras yra tas pats:



7 pav. Dvi stebinių aibės gali turėti tas pačias centro pozicijas, bet skirtingą sklaidos apie šį centrą pavidalą

3. Panagrinėkime šį pavyzdį detaliau. Taškų $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 5$, $x_5 = 7$, $x_6 = 9$, $x_7 = 12$, $x_8 = 13$ ir $x_9 = 14$ imties vidurkis $\bar{x} = (0 + 1 + 2 + 5 + 7 + 9 + 12 + 13 + 14)/9 = 7$, o y -taškų imties vidurkis $\bar{y} = (2 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 12)/9 = 7$ yra toks pats. Apibrėžus i -tojo stebinio nuokrypį formule

Nuokrypis $\Delta_i = i$ -tasis stebinys – (Imties vidurkis),

gauname dvi nuokrypių sekas: 7, -6, 5, -2, 0, 2, 5, 6, 7 bei -5, -3, -2, -1, 0, 1, 2, 3, 5. Susumavę šiuos nuokrypius, abiem atvejais gauname vieną ir tą patį rezultatą:

$$\sum \Delta_i = 0.$$

Matome, kad teigiamus nuokrypius „atsveria“ neigiami nuokryptai, todėl ir bendru atveju galima įrodyti, kad visų nuokrypių Δ_i suma visada lygi nuliui.

4. Todėl, norėdami gauti sklaidos matą, turime pašalinti neigiamų nuokrypių Δ_i ženklus, prieš juos sudedant. Vienas iš tokių ženklų pašalinimo būdų –

pakelti nuokrypius kvadratu. Taigi, pakėlus nuokrypius kvadratu, juos susumavus ir padauginus iš normavimo koeficiento $1/(n-1)$ (dažnai normavimo koeficientu pasirenkamas $1/n$) gaunamas **sklaidos ir kitimo matas**, vadinamas *imties dispersija*¹ arba *empirine dispersija*:

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

Statistikoje naudinga ir *empirinio standartinio nuokrypio* sąvoka:

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

5. Apskaičiuosime empirinę dispersiją ir empirinį standartinį nuokrypį 2-ame punkte pateiktame pavyzdyje. Tuo tikslu sudarome 7 lentelę.

7 lentelė

Stebiniai x_i	Nuokrypiai $x_i - \bar{x}$	(Nuokrypiai) ² $(x_i - \bar{x})^2$	Stebiniai y_i	Nuokrypiai $y_i - \bar{y}$	(Nuokrypiai) ² $(y_i - \bar{y})^2$
0	-7	49	2	-5	25
1	-6	36	4	-3	9
2	-5	25	5	-2	4
5	-2	4	6	-1	1
7	0	0	7	0	0
9	2	4	8	1	1
12	5	25	9	2	4
13	6	36	10	3	9
14	7	49	12	5	25
Iš viso	0	228	Iš viso	0	78

Kadangi $n-1=8$, tai iš lentelės gauname, kad

$$s_x^2 = 228/8 = 28,5, \quad s_x = \sqrt{28,5} = 5,34,$$

$$s_y^2 = 78/8 = 9,75, \quad s_y = \sqrt{9,75} = 3,12.$$

¹ Angl. *sample variance*, rus. *выборочная дисперсия*.

6. Perėjimas nuo imties dispersijos apibrėžimo prie atsitiktinio dydžio dispersijos apibrėžimo analogiškas perėjimui nuo imties vidurkio apibrėžimo prie atsitiktinio dydžio vidurkio apibrėžimo, todėl mes jo detaliau nekartosime.

Kadangi vidurkis μ yra atsitiktinio dydžio X centras, mes išreikšime jo sklaidą nuokrypio $(X - \mu)$ terminais. O būtent, *atsitiktinio dydžio X dispersiją* žymėsime DX ir apibrėšime kaip nuokrypio kvadrato $(X - \mu)^2$ matematinę viltį. Dydžio $(X - \mu)^2$ matematinė viltis gaunama kiekvieną reikšmę $(x_i - \mu)^2$ padauginus iš tikimybės $f(x_i)$ ir po to sumuojant visas sandaugas:

$$\begin{aligned} DX &= \sum (\text{nuokrypis})^2 \times (\text{tikimybė}) = \\ &= \sum (x_i - \mu)^2 f(x_i). \end{aligned}$$

Dydis $\sigma = \sqrt{DX}$ vadinamas atsitiktinio dydžio X *standartiniu nuokrypiu*. Atsitiktinio dydžio X dispersija taip pat vadinama *populiacijos dispersija*, o σ vadinamas *populiacijos standartiniu nuokrypiu*.

7. Kai tyrinėtojas analizuoja statistinius duomenis, jis paprastai remiasi dviem charakteristikomis – empiriniu vidurkiu $\bar{x}$ ir empiriniu standartiniu nuokrypiu s . Rusų matematikas P.L. Čebyševas (1821–1894) įrodė, kad minimalus procentas stebinių, kurie patenka į intervalą $(\bar{x} - ks, \bar{x} + ks)$, kur k – natūrinis skaičius, yra $(1 - 1/k^2) \cdot 100$. Detalizuosime šį teiginį labiausiai paplitusiems praktikoje atvejams $k = 2$ ir $k = 3$.

Čebyševio taisyklė. *Mažiausiai 75% stebinių patenka į intervalą $(\bar{x} - 2s, \bar{x} + 2s)$. Mažiausiai 89% stebinių patenka į intervalą $(\bar{x} - 3s, \bar{x} + 3s)$.*

Turint papildomą informaciją apie stebinius (pvz., žinant, kad jų dažnio funkcija yra simetrinė), Čebyševio taisyklę galima sustiprinti. O būtent, tokiu atveju ne mažiau kaip 95% stebinių patenka į intervalą $(\bar{x} - 2s, \bar{x} + 2s)$ ir ne mažiau kaip 99%, t.y. praktiškai visi stebiniai – į intervalą $(\bar{x} - 3s, \bar{x} + 3s)$.

8. TIKIMYBINIO TANKIO KREIVĖ KAIP HISTOGRAMŲ RIBA. TANKIO SAVYBĖS

1. Penktame skyrelyje mes matėme, kad *diskretusis atsitiktinis dydis* – tai dydis, kuris gali įgyti tik izoliuotas reikšmes, kurios dažniausiai atspindi skaičiavimą. Dabar sukoncentruosime savo dėmesį į *tolydžiuosius atsitiktinius dydžius*, kurie gali įgyti bet kokią reikšmę iš intervalo. Tai dydžiai, įgyjantys reikšmes iš tolydžios skalės, tokie kaip svoris, ilgis, gyvenimo trukmė ir pan.

2. Kadangi tikimybę suvokiame kaip santykinio dažnio statistinio stabilumo fenomeną, tai tolydaus atsitiktinio dydžio tankio sąvoką gausime kaip santykinio dažnio tankių histogramų¹, sudarytų didinant matavimų skaičių ir tuo pačiu siaurinant klasių plotį, ribą.

3. Šią schemą iliustruosime konkrečiu pavyzdžiu² apie naujagimių svorį X . Tarkime, kad užregistruoti 100 naujagimių svoriai, gauti duomenys sugrupuoti į klases po 400 g, o atitinkami santykiniai dažniai pavaizduoti 8 pav. histograma (A). Priminsime, kad *santykinio dažnio histograma turi šias savybes*:

- 1) plotas, kurį riboja histograma, lygus 1;
- 2) jei a ir b – bet kurie du klasių ribiniai taškai, tai santykinio dažnio reikšmė intervale $[a; b]$ lygi sumai plotų, kuriuos histograma riboja intervale $[a; b]$.

Pavyzdžiui, 8 pav. (A) parinkus $a = 3$, $b = 3,8$ gauname, kad santykinio dažnio reikšmė šiame intervale yra $0,28 + 0,25 = 0,53$, t.y. iš 100 naujagimių net 53-ų svoris yra intervale nuo 3 iki 3,8 kg.

4. Tarkime toliau, kad stebinių, t.y. naujagimių, kurių svoris stebimas, skaičius išaugo iki 5 000, o gauti duomenys sugrupuoti į klases po 100 g. Santykinis dažnis šiam atvejui pavaizduotas 8 pav. histograma (B). Ji yra histogramos (A) patikslinimas, nes sudaryta didesnio stebinių skaičiaus pagrindu ir atspindi santykinį dažnį siauresniems klasių intervalams.

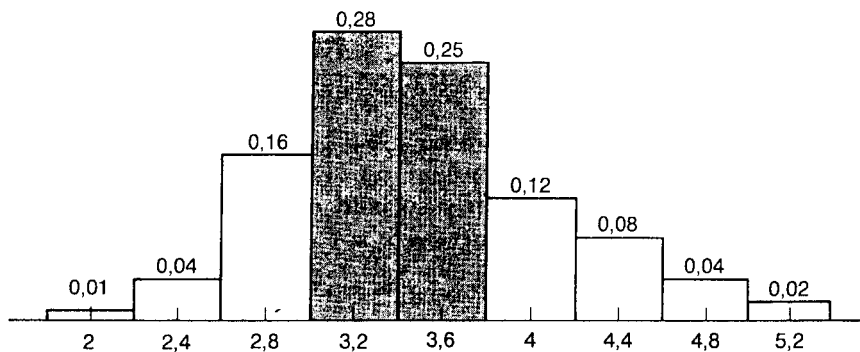
5. Tokiu pat būdu tęskime šį procesą, dabar mintyse įsivaizduodami, kaip kis histograma, didėjant stebinių skaičiui ir siaurėjant klasių intervalams. Šio histogramų kitimo riba pavaizduota 8 pav. (C).

Kadangi *tikimybė yra interpretuojama kaip santykinio dažnio statistinio stabilumo fenomenas*, tai kreivė, kuri gaunama kaip santykinio dažnio histogramų ribinė forma, atspindi būdą, kuriuo visa tikimybė masė 1 pasiskirsčiusi virš visų atsitiktinio dydžio X galimų reikšmių intervalo. Ši kreivė vadinama atsitiktinio dydžio X *tikimybinių tankio kreive*. Matematinė funkcija $f(x)$, kurios grafikas yra ši tankio kreivė, vadinama atsitiktinio dydžio X *tankio funkcija*, arba tiesiog *tankiu*.

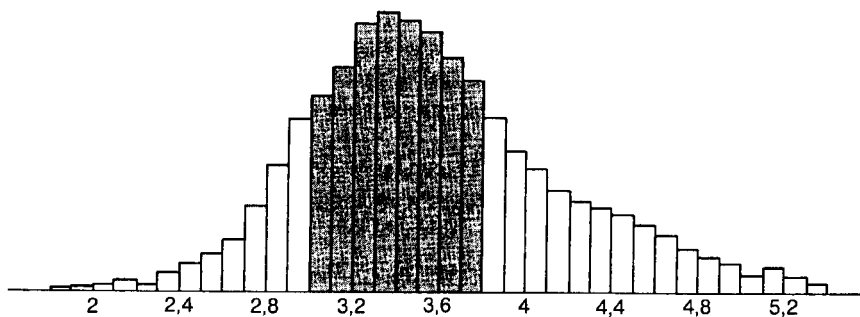
6. Histogramos savybės 1) ir 2), kurias mes suformulavome punkte 3, po ribinio perėjimo išlieka galioti ir tikimybinių tankio kreivei. Be to, kadangi

¹ Toliau šias histogramas sutrumpintai vadinsime *santykinio dažnio histogramomis*.

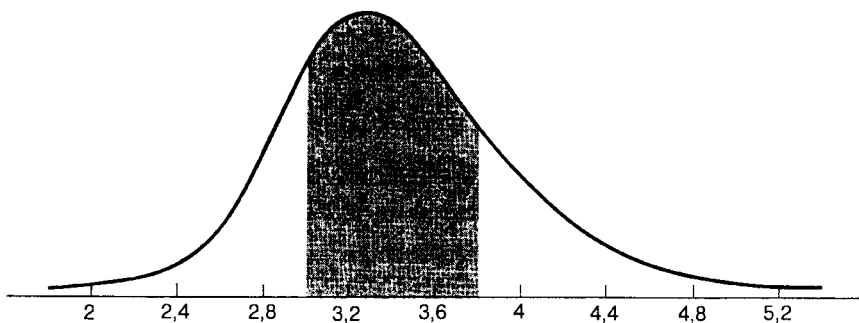
² Pagal R.A. Johnson, G.K. Bhattacharyya [5] autorinį pavyzdį.



(A) 100 naujagimių svorių santykinų dažnių histograma su klasių pločiais 0,4 kg



(B) 5000 naujagimių svorių santykinų dažnių histograma su klasių pločiais 0,1 kg



(C) Tolygaus atsitiktinio dydžio X (= naujagimio svoris) tikimybinių tankio funkcija

8 pav. Tikimybinių tankio kreivė, pavaizduota kaip santykinio dažnio histogramų ribinė forma

histograma negali būti žemiau x ašies, ta pati pastaba galioja ir tikimybinio tankio kreivei. Todėl *tikimybinio tankio funkcija* $f(x)$ turi šias savybes:

- 1) bendras plotas po tikimybinio tankio $f(x)$ kreive lygus 1;
- 2) tikimybė $P(a \leq X \leq b)$ lygi plotui po tikimybinio tankio $f(x)$ kreive tarp a ir b ;

- 3) $f(x) \geq 0$ visiems $x \geq 0$.

7. Jei $a = b$, tai iš 2) savybės gauname, kad

$$P(X = x) = 0 \quad \text{bet kuriam } x,$$

nes plotas figūros, sukoncentruotos viename taške, lygus nuliui.

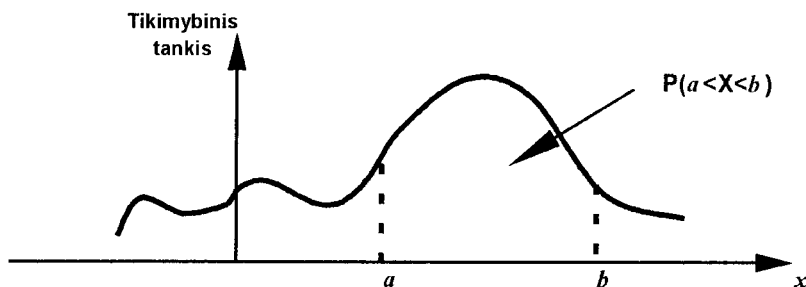
Taigi, skirtingai nuo tikimybės masės funkcijos diskrečiuoju atveju, tikimybiniis tankis tolydžiuoju atveju nereprezentuoja įvykio $\{X = x\}$. Jei X – tolydus atsitiktinis dydis, tai įvykio $\{X = x\}$ tikimybė visada lygi nuliui. Yra prasmė kalbėti tik apie tikimybę įvykio, kad X yra kažkokiame intervale.

8. Ypatingą svarbą tikimybių teorijoje ir matematinėje statistikoje turi intervalai $(-\infty; x)$, o funkcija $F(x) = P(X < x)$ vadinama statistinio dydžio X **skirstiniu**¹ arba, pagal Lietuvos standartą, **skirstinio funkcija**.

9. Kadangi tolydaus atsitiktinio dydžio X atveju $P(X = x) = 0$ bet kuriam realiam x , tai

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b).$$

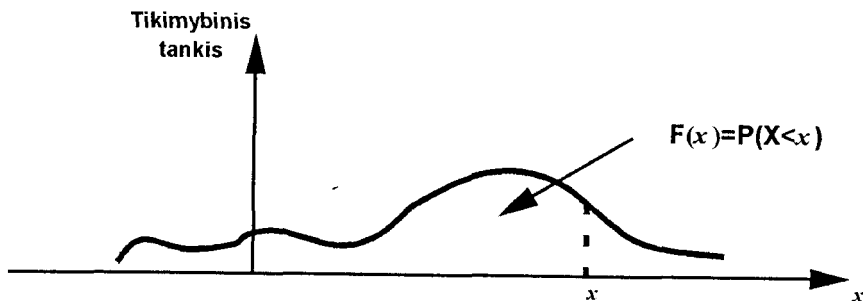
Ši tikimybė, kaip seka iš 2) tankio funkcijos savybės, lygi plotui po atsitiktinio dydžio X tikimybinio tankio kreive tarp a ir b (žr. 9 pav.).



9 pav. Plotas po atsitiktinio dydžio X tikimybinio tankio kreive tarp a ir b lygus tikimybei $P(a < X < b)$

¹ Angl. *distribution function*, rus. *функция распределения*.

10. Paėmę $a = -\infty$, $b = x$, iš čia nesunkiai gauname, kad skirstinio $F(x)$ reikšmė taške x lygi plotui po tikimybinių tankio kreive iki taško x , t.y. plotui po tankio kreive pusašėje $(-\infty; x)$ (žr. 10 pav.).



10 pav. Skirstinio reikšmė taške x lygi plotui po tikimybinių tankio kreive iki taško x

9. TOLYGUSIS SKIRSTINYS. MEDIANA, KVANTILIS, PROCENTILIS, KVARTILIS

1. Skyrelį pradėsime pastaba, kad žodžiai *tolygusis* ir *tolydusis* yra ne tik skirtingi, bet ir išreiškia skirtingas sąvokas. Mes nagrinėjame *tolydžiuosius atsitiktinius dydžius*, kurie gali įgyti bet kokią reikšmę iš tam tikro intervalo – baigtinio ar begalinio. Tolygusis skirstinys yra tolydžiojo atsitiktinio dydžio skirstinio atskiras atvejis, nors pagal Lietuvos standarto pateikiamą apibrėžimą juo gali būti ir diskrečiojo atsitiktinio dydžio skirstinys.

2. *Tolygusis skirstinys*¹ – tai

- 1) tolydžiojo atsitiktinio dydžio, kurio tikimybinių tankis yra pastovus baigtiniame intervale $[a; b]$ ir lygus nuliui už šio intervalo, skirstinys.
- 2) diskrečiojo atsitiktinio dydžio, kurio tikimybinių masės funkcija yra

$$P(X = x_i) = \frac{1}{n} \quad (i = 1, 2, \dots, n),$$

o taškai x_i išsidėstę vienodais atstumais, skirstinys.

¹ Angl. *uniform distribution, rectangular distributions*; rus. *равномерное распределение, прямоугольное распределение*.

3. Taigi, praeitame skyrelyje išnagrinėjome tolydžiojo atsitiktinio dydžio sąvoką, dabar išnagrinėsime šio dydžio skirstinio konkretų pavyzdį – tolygųjį skirstinį. Tolimesniam nagrinėjimui yra tikslinga prisiminti *klasės santykinio dažnio tankio* sąvoką, kuri buvo įvesta 3-ajame skyrelyje:

$$\left(\begin{array}{c} \text{Klasės santykinio} \\ \text{dažnio tankis} \end{array} \right) = \frac{\text{Klasės santykinis dažnis}}{\text{Klasės plotis}}$$

Ši sąvoka, kaip matyti iš žemiau nagrinėjamo pavyzdžio, ypač patogi, norint santykinio dažnio tankio histograma pavaizduoti klasės santykinį dažnį:

Histogramos stačiakampio plotas =

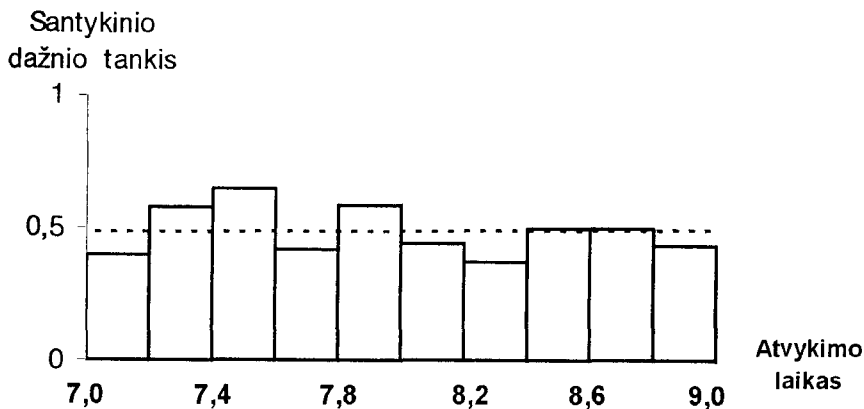
$$= (\text{Pagrindo plotis}) \times (\text{Stačiakampio aukštis}) =$$

$$= (\text{Klasės plotis}) \times (\text{Klasės santykinio dažnio tankis}) =$$

$$= (\text{Klasės plotis}) \times \frac{\text{Klasės santykinis dažnis}}{\text{Klasės plotis}} =$$

$$= \text{Klasės santykinis dažnis}$$

Pavyzdys. Maisto prekių parduotuvė pradeda darbą 9 val. ryto. Jos darbuotojas atvyksta į darbą tarp 7 ir 9 val. Susumavus jo atvykimo į darbą duomenis metų bėgyje, gauta tokia santykinio dažnio histograma:



II pav. Darbuotojo atvykimo į darbą santykinio dažnio histograma

Tegu X yra atsitiktinis dydis, išreiškiantis darbuotojo atvykimo į parduotuvę laiką.

(A) Rasti X tankio kreivę kaip histogramų, gaunamų didinant stebinių skaičių ir mažinant atvykimo laiko intervalų plotį, ribą.

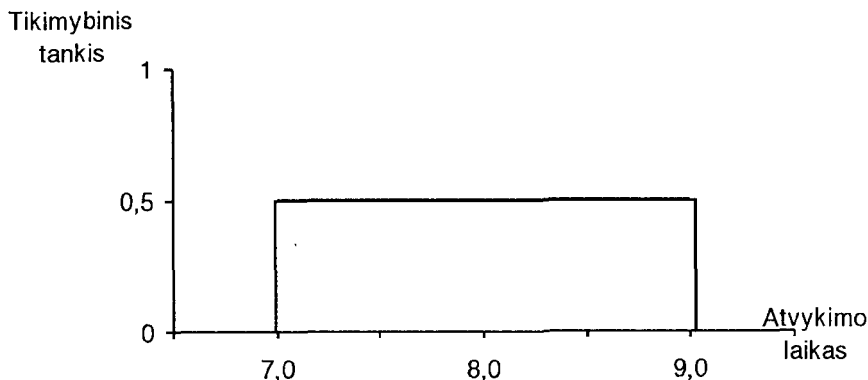
(B) Apskaičiuoti tikimybę įvykio, kad darbuotojas atvyks į parduotuvę tarp pusės aštuonių ir 15 min. po aštuonių.

Sprendimas. Pirmiausia naudinga suprasti, kokią informaciją teikia pavyzdžio histograma. Tarkime, kad metų bėgyje darbuotojas atvyko į darbą 260 kartų. Pirmojo stačiakampio (t.y. stulpelio) plotis yra $7,2 - 7,0 = 0,2$, o aukštis, kaip matyti iš histogramos, 0,442. Taigi, pirmojo stačiakampio plotas yra 0,088, o tai, kaip ką tik matėme, yra klasės santykinis dažnis. Kadangi

$$\text{Klasės santykinis dažnis} = \frac{\text{Klasės elementų skaičius}}{\text{Visų stebinių skaičius}}$$

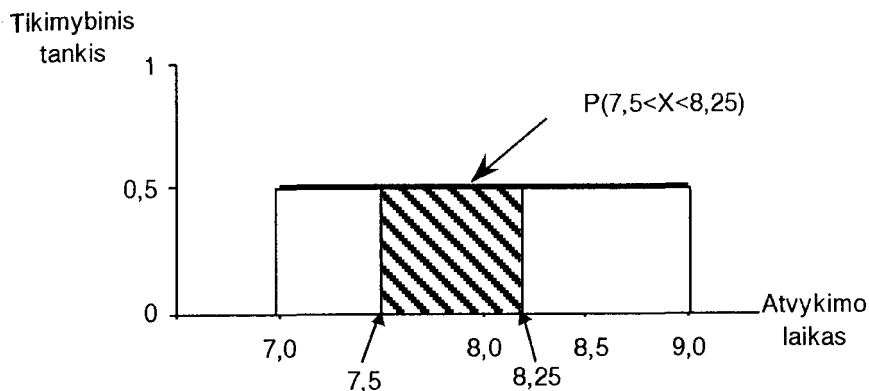
tai tarp 7 val. 00 min. ir 7 val. 12 min. darbuotojas į darbą atvyko $0,88 \times 260 \approx 23$ kartus.

Kadangi visų histogramos stačiakampių plotų suma lygi 1, o šių stačiakampių aukštis svyruoja apie 0,5, tai kreivė, kuri gaunama kaip santykinio dažnio histogramų riba, yra tokia:



12 pav. Tolygusis tikimybiniis tankis kaip santykinų dažnių histogramų riba

Dabar apskaičiuosime tikimybę įvykio, kad darbuotojas atvyks į parduotuvę tarp 7 val. 30 min. ir 8 val. 15 min. Kadangi 30 ir 15 minučių sudaro atitinkamai 0,5 ir 0,25 valandos dalį, tai mums reikia rasti užbrūkšniuoto stačiakampio, kuris yra pavaizduotas 13 pav. tarp 7,5 ir 8,25, plotą.



13 pav. Konkretios tikimybės apskaičiavimas pagal tolygiojo tikimybinių tankio grafiką

Kadangi užbrūkšniuotojo stačiakampio plotis yra $8,25 - 7,5 = 0,75$, o aukštis yra 0,5, tai

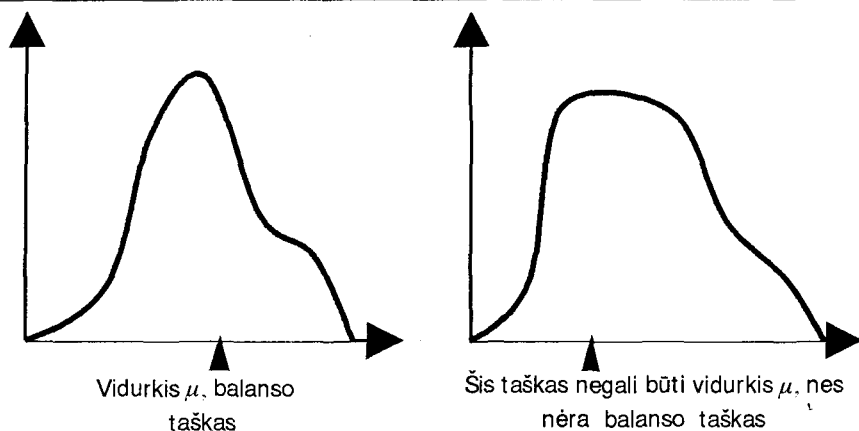
$$P(7,5 < X < 8,25) = 0,75 \times 0,5 = 0,375.$$

4. Tačiau daugeliu atvejų, kai tikimybinių tankio kreivė yra sudėtingesnė, rasti tokio tipo plotą be integralinio skaičiavimo metodų tiesiog neįmanoma. Tokiais atvejais būna labai naudingos lentelės. Be to, dažnai pakanka atsitiktinių dydžių skaitinių charakteristikų – tokių kaip vidurkis, mediana, kvantilis, procentilis ir kt.

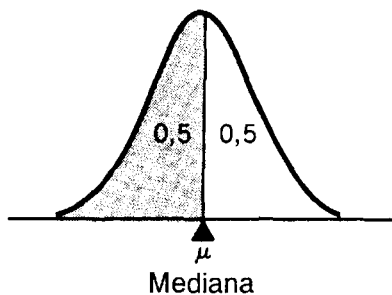
5. Panašiai kaip ir diskrečiuoju atveju, fizikinė vidurkio interpretacija galėtų būti tokia. Plotą, kurį riboja tikimybinių tankis, įsivaizduokime kaip figūrą iš plonos vienarūšės medžiagos. Šios medžiagos balanso taškas μ išilgai horizontalios ašies ir bus vidurkis (žr. 14 pav.).

Sklaidą apie vidurkį charakterizuoja standartinis nuokrypis. Juo standartinis nuokrypis σ yra mažesnis, juo aukštesnė ir siauresnė tikimybinių tankio kreivė. Ir atvirkščiai, juo standartinis nuokrypis σ yra didesnis, juo ši kreivė žemesnė ir plokštesnė (žr. 14 pav.).

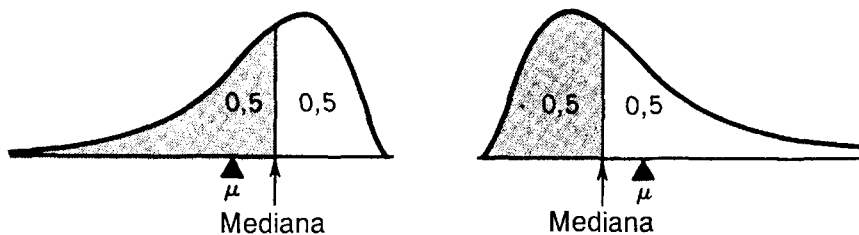
6. Sekančiame, t.y. 15-ajame paveiksle vidurkis μ taip pat pavaizduotas kaip balanso taškas. Čia taip pat pavaizduota dar viena skaitinė charakteristika, dar vienas centro matas – mediana. Tai atsitiktinio dydžio X reikšmė, dalijanti plotą po tikimybinių tankio kreivę į dvi lygias dalis.



14 pav. Fizikinė vidurkio interpretacija tais atvejais, kai egzistuoja tankis



15 pav. Tankio su vienodais vidurkiu bei mediana pavyzdys



16 pav. Tankių su nevienodais vidurkiu bei mediana pavyzdžiai

Jei tikimybinių tankio kreivė yra simetriška, tai vidurkis ir mediana sutampa. Priešingu atveju, kaip matome iš 16 pav., taip gali ir nebūti.

7. Pagal Lietuvos standartą, (atsitiktinio dydžio arba skirstinio) *p*-kvantilis¹ yra atsitiktinio dydžio reikšmė, su kuria skirstinio funkcija arba lygi p ($0 \leq p \leq 1$), arba peršoka nuo mažesnės už p reikšmės prie didesnės už p reikšmės.

Standarto pastabos:

1) Jei skirstinio funkcija lygi p visame intervale tarp dviejų gretimų atsitiktinio dydžio reikšmių, tai kiekvieną šio intervalo reikšmę laikome p -kvantiliu.

2) x_p yra p -kvantilis, jei

$$P(X < x_p) \leq p \leq P(X \leq x_p).$$

3) Tolydžiojo atsitiktinio dydžio p -kvantilis yra atsitiktinio dydžio reikšmė, žemiau kurios yra skirstinio p -toji dalis.

4) Procentilis apibrėžiamas panašiai, tik p išreiškiamas procentais.

8. Pagaliau, **mediana** – tai 0,5-kvantilis, o **kvartilis** – tai 0,25-kvantilis arba 0,75-kvantilis.

10. NORMALUSIS SKIRSTINYS IR JO TANKIO SAVYBĖS

Seniai buvo pastebėta, kad glodi varpo formos kreivė, kuria išreiškiamas tikimybinis tankis, tinka įvairių matematinių modelių aprašymui. Tokia varpo formos kreivė vadinama *normaliąja kreive*, o ją atitinkantis tikimybinis tankis – *normaliuoju tankiu*. Matematiškai jis apibrėžiamas formule

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty,$$

kur μ yra populiacijos vidurkis, o σ – populiacijos standartinis nuokrypis.

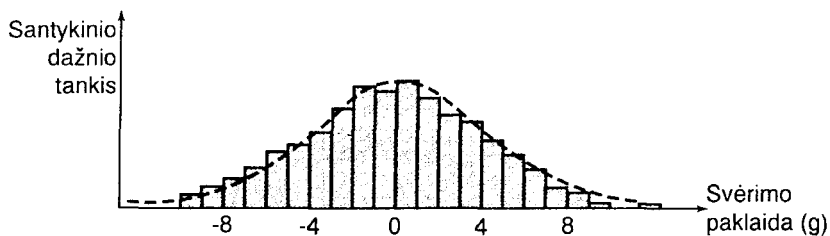
Atitinkantis šį tankį *normalusis skirstinys*² – plačiausiai naudojamas tolydusis skirstinys, kuriam tenka pagrindinis vaidmuo statistinėse išvadose.

Išnagrinėsime keletą praktinių pavyzdžių.

2. Tarkime, kad X yra atsitiktinis dydis, reiškiantis paklaidą, kuri daroma serijiniu būdu fasuojant cukrų į maišelius po 1 kg. Tarsime, kad paklaida yra teigiama, jei maišelis perpildomas, ir neigiama, jei yra cukraus trūkumas. Atlikus kruopštų 1000 maišelių kontrolinį svėrimą, buvo gauta tokia santykinio dažnio histograma:

¹ Angl. *quantile, fractile*; rus. *квантиль*.

² Angl. *normal distribution*; *Laplace-Gauss distribution*.

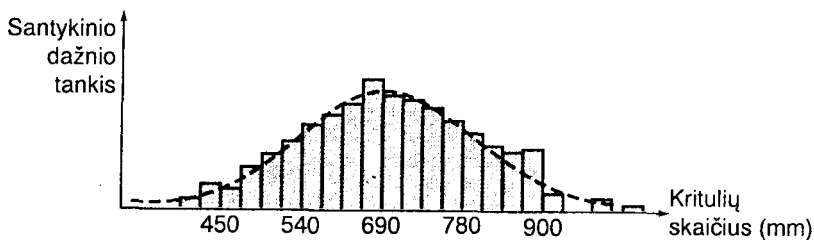


17 pav. Cukraus fasavimo paklaidos histograma

Suprantama, mes tikimės, kad maišeliuose cukraus nei per daug, nei per mažai, t.y. svėrimo paklaida yra artima nuliui. Tai patvirtina gautoji histograma. Didėjant pertekliaus ar trūkumo reikšmėms, vis mažesnis ir mažesnis yra tokių maišelių kiekis.

Didinant kontrolinių maišelių kiekį nuo 1000 iki 2000, 3000 ir t.t., galėtume kalbėti apie histogramų ribą. 17 pav. ši riba pavaizduota kaip normaliojo tankio kreivė (punktyrinė linija).

3. Kitas pavyzdys galėtų būti kritulių kiekio per metus, išskrintančių, tarkime, Vilniuje, histograma:

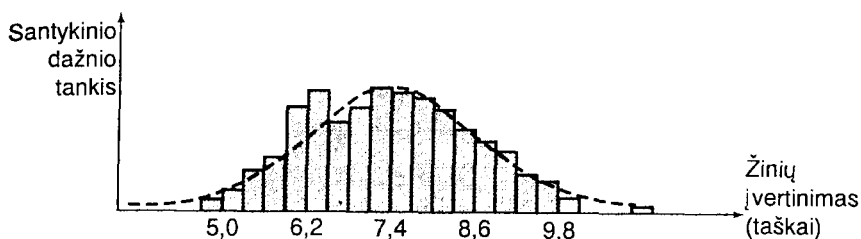


18 pav. Kritulių skaičiaus per metus histograma

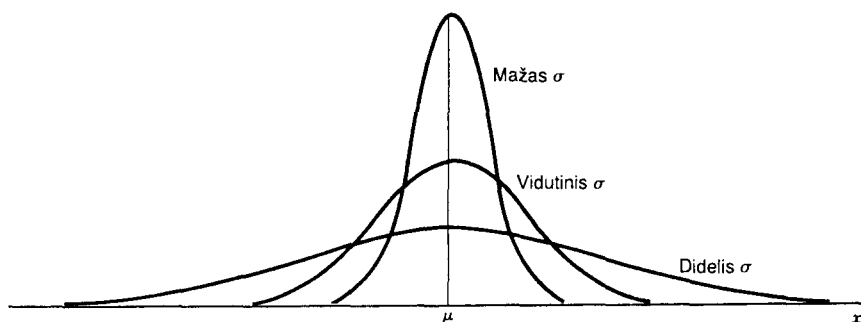
Analogiškai sudaroma studentų, laikiusių statistikos egzaminą, žinių įvertinimo histograma (žr. 19 pav.).

4. Normaliojo tankio kreivė (priminsime, kad sutrumpintai ją pavadiname normaliaja kreive) yra simetriška vidurkio μ atžvilgiu, jame yra sukoncentruota varpo viršūnė. Skirtingos standartinio nuokrypio reikšmės sąlygoja skirtingą varpo aukštį ir formą (žr. 20 pav.).

Taigi, mažėjant standartiniam nuokrypiui σ , didėja varpo aukštis, o jo masė labiau koncentruojasi apie vidurkį μ .



19 pav. Studentų žinių įvertinimų histograma

20 pav. Normaliojo tankio kreivės skirtingoms σ reikšmėms. Šios kreivės yra simetriškos μ atžvilgiu

5. Normalusis skirstinys, kurio vidurkis μ lygus nuliui, o standartinis nuokrypis σ lygus 1, vadinamas *standartiniu normaliuoju skirstiniu*, o pats dydis – *standartiniu normaliuoju atsitiktiniu dydžiu* ir vakarietiškuose vadovėliuose paprastai žymimas raide Z . Standartinė normalioji kreivė pavaizduota 21 pav.

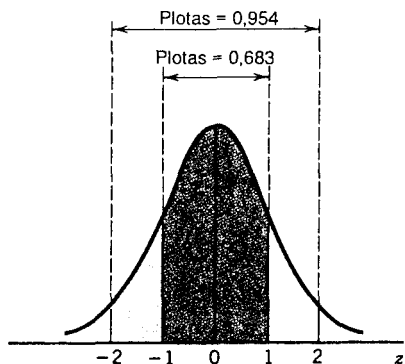
Kadangi standartinė normalioji kreivė yra simetriška nulinio atžvilgiu, tai

- 1) $P(Z < 0) = 0,5$;
- 2) $P(Z < -z) = 1 - P(Z \leq z) = P(Z > z)$.

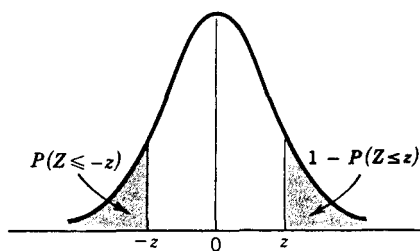
Grafiškai tai galima pavaizduoti taip, kaip parodyta 22 pav.

6. Galima parodyti, kad jei N yra normalusis atsitiktinis dydis su vidurkiu μ ir standartiniu nuokrypiu σ , tai

$$P(a < N < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right).$$



21 pav. Standartinė normalioji kreivė.
95,4% ploto, esančio po šia kreive,
sukoncentruota intervale $[-2;2]$.



22 pav. Normaliųjų tankių uodegos yra
lygios

Toks perėjimas nuo N prie Z vadinamas *standartizacija* ir yra ypač naudingas, norint pasinaudoti lentelėmis.

Pailiustruosime tai, apskaičiuodami tikimybes $P(\mu - \sigma < N < \mu + \sigma)$, $P(\mu - 2\sigma < N < \mu + 2\sigma)$ ir $P(\mu - 3\sigma < N < \mu + 3\sigma)$. Pastebėsime, kad bet kuriam k

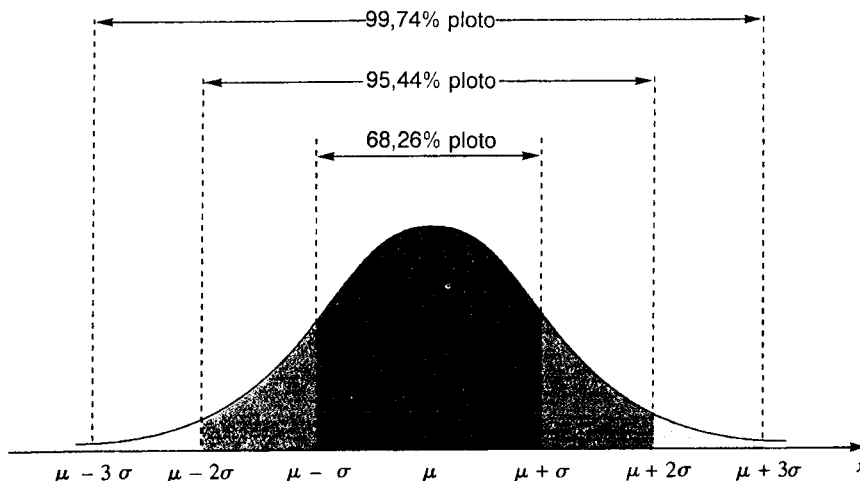
$$\begin{aligned} P(\mu - k\sigma < N < \mu + k\sigma) &= P\left(\frac{(\mu - k\sigma) - \mu}{\sigma} < Z < \frac{(\mu + k\sigma) - \mu}{\sigma}\right) = \\ &= P(-k < Z < k). \end{aligned}$$

Dabar belieka pasinaudoti lentelėmis (žr. Priedą vadovėlio gale) ir įsitikinti, kad

$$P(\mu - \sigma < N < \mu + \sigma) = P(-1 < Z < 1) = 2 \times 0,3413 = 0,6826;$$

$$P(\mu - 2\sigma < N < \mu + 2\sigma) = P(-2 < Z < 2) = 2 \times 0,4772 = 0,9544;$$

$$P(\mu - 3\sigma < N < \mu + 3\sigma) = P(-3 < Z < 3) = 2 \times 0,4987 = 0,9974.$$



23 pav. Grafinis ir procentinis įvykių $\{\mu - k\sigma < N < \mu + k\sigma\}$ tikimybių vaizdavimas

Grafiškai tai galima būtų pavaizduoti taip, kaip parodyta 21 pav.

7. Teoriškai normalioji kreivė yra neaprežta abiejomis kryptimis, o normalusis atsitiktinis dydis gali įgyti bet kokią reikšmę nuo minus begalybės iki plus begalybės. Praktikoje tačiau gi, kaip matyti iš 23 pav., šį atsitiktinį dydį pakanka nagrinėti intervale $(\mu - 3\sigma, \mu + 3\sigma)$, nes jame sukoncentruota net 99,74% viso ploto, esančio po normaliaja kreive.

8. Normaliajam skirstiniui ypatingas vaidmuo tenka taip vadinamoje *centrinėje ribinėje teoremoje*. Ši teorema tvirtina, kad, esant pakankamai bendroms prielaidoms, didelio skaičiaus atsitiktinių dydžių sumos skirstinys yra artimas normaliajam skirstiniui.

11. STATISTINIŲ IŠVADŲ SUPRATIMAS PRAKTINIŲ MODELIŲ KONTEKSTE

1. *Statistinės išvados*¹ – tai statistikos pagrindas. Jos leidžia tyrinėtojiui pagrįsti perėjimą nuo atskiros imties stebinių prie bendro atvejo. Šiame skyrelyje detaliau panagrinėsime šią koncepciją, prieš tai įvesdami kelias bazines sąvokas.

¹ Angl. *statistical inference*; rus. *статистические выводы*.

2. Tipiškas statistikos uždavinys – tai populiacijos skaitinių charakteristikų analizė. Populiacijos skaitinių charakteristikų pavyzdžiai – tai visumos, turinčios tam tikrą savybę, proporcijos visoje populiacijoje apskaičiavimas, populiacijos vidurkio arba populiacijos standartinio nuokrypio radimas ir pan. *Skaitinė populiacijos charakteristika vadinama parametru.*

3. Jei turime tik populiacijos imtį, tai mūsų statistinės išvados apie populiacijos parametą, suprantama, turi būti grindžiamos tik šios imties stebiniais. Šių stebinių funkcija dar yra vadinamas *statistika*¹.

Pavyzdžiui, *empirinis vidurkis* $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ yra imties $X_1, \dots, X_n$ funkcija, o tuo pačiu ir statistika. Beje, tikimybių teorijos požiūriu, *imtis* $X_1, \dots, X_n$ – tai seka nepriklausomų atsitiktinių dydžių, turinčių tą patį skirstinį.

4. Taigi, tikroji populiacijos parametro reikšmė yra nežinoma konstanta. Beje, ši konstanta gali būti tiksliai apskaičiuota tik pilnai ir visapusiškai išnagrinėjus populiaciją. Tačiau tai, kaip pamatysime iš konkretaus pavyzdžio, labai dažnai yra arba neįmanoma, arba praktiškai netikslinga (pavyzdžiui, pternelyg brangu). Tokiais atvejais itin išryškėja *statistinių išvadų vaidmuo*.

Išnagrinėsime konkretų pavyzdį.

5. Tarkime, kad mus domina toks uždavinys. Nustatyti, kokia dalis Lietuvos gyventojų bent kartą per metus perka bent vieną lietuviškų batų porą. Tiksliai šios proporcijos skaitinė reikšmė negali būti žinoma tol, kol nebus apklaustas kiekvienas Lietuvos gyventojas. Žinoma, tokia apklausa avalynės gamintojams yra pernelyg brangi.

Todėl, prisiminus statistinio stabilumo fenomeną, gana logiškai atrodo tokia prielaida. Apklausus, pavyzdžiui, 100 gyventojų ir tokiu būdu radus naudojančių lietuvišką avalynę (šioje atsitiktinai pasirinktų gyventojų grupėje) dalies skaitinę reikšmę skaityti, kad ši reikšmė yra artima ieškomajai. Tai yra, logiška skaityti, kad santykis gyventojų, naudojančių lietuvišką avalynę, pasirinktoje imtyje yra artimas analogiškam santykiui visų Lietuvos gyventojų tarpe. Žinoma, labai svarbu, kad pasirinkta imtis tikrai būtų atsitiktinė, bet tai jau kita šneka.

Gali atsitikti taip, kad, apklausus 100 gyventojų, išaiškėjo, jog lietuvišką avalynę naudoja 43 apklaustieji, todėl ieškomosios proporcijos skaitinė reikšmė pasirenkame 0,43. Apklausus kitą šimtą gyventojų galime gauti, kad lietuvišką

¹ Ši sąvoka bus patikslinta II dalyje – ne bet kokia stebinių funkcija yra vadinama statistika.

avalynę perka 34 apklaustieji, t.y. imties proporcijos reikšmė yra 0,34.

6. Kiekvieną kartą apklausus naują 100 žmonių grupę, t.y. gavus naują nagrinėjamos imties realizaciją, gauname kitą imties proporcijos reikšmę. Pažymėjus imties proporciją raide X ir nagrinėjant ją kaip atsitiktinį dydį, būtų labai pageidautina rasti jos skirstinį.

Šio tikslo prasmę nesunku suvokti. Iš pirmosios imties padarėme išvadą, kad imties proporcija X įgijo reikšmę $x = 0,43$. Ar reikšmė 0,43 pilnai atspindi imties proporciją, t.y. ar šis skaičius ir yra ieškomoji dalis visų Lietuvos gyventojų, naudojančių lietuvišką avalynę? Gal taip, o gal ir ne.

7. Tačiau jau prieš pasirinkdami imtį mes intuityviai jautėme, kad labai tikėtina, jog proporcija pasirinktoje gyventojų imtyje yra artima proporcijai visų gyventojų tarpe. O ką reiškia žodžių junginys „labai tikėtina“? Kalbant tiksliau, kokia tikimybė, kad proporcija pasirinktoje gyventojų imtyje (sutrumpintai – imties proporcija) skiriasi nuo ieškomosios, tarkime, ne daugiau kaip 0,05?

Mes galime atsakyti į šį klausimą, o taip pat į šios rūšies klausimus, jei žinome imties proporcijos X skirstinį. Analogiškai, naudinga žinoti kitų statistikų (pavyzdžiui, empirinio vidurkio $\bar{X}$ ir empirinės dispersijos S^2) skirstinius.

8. Žinoma, skirstinio radimas yra matematinis, t.y. teorinis uždavinys, kurį praktinių modelių kontekste galima būtų aprašyti taip.

Išsivaizduokime visas galimas didumo n imtis, kurios gali būti paimtos iš nagrinėjamos populiacijos. Kiekvienai iš šių imčių galime apskaičiuoti mus dominantį santykinį dažnį. Didindami imtį, o tuo pačiu ir skaičių n , kuriuo išreiškiamas imties dydis, ir kaskart apskaičiuodami santykinio dažnio tankį bei konstruodami atitinkamas histogramas, mes galime gauti pakankamai tikslią atsitiktinio dydžio X teorinio skirstinio aproksimaciją.

9. Trumpai reziumuosime pagrindinę mintį. Norime sužinoti tikrąją arba bent apytikslę populiacijos parametro reikšmę. Turime tik šios populiacijos imtį. Šios imties pagrindu konstruojame statistiką (pavyzdžiui, empirinį vidurkį $\bar{X}$). Turint konkrečią imtį, turime ir konkrečią statistikos reikšmę (pavyzdžiui, empirinio vidurkio reikšmę $\bar{x}$). Ar ši konkreti reikšmė yra artima ieškomajam populiacijos parametrai?

Į šį ir panašius klausimus turėtume atsakymą, jei žinotume *statistikos skirstinį*. Tačiau tai yra sudėtingas uždavinys. Todėl sekančiuose trijuose skyreliuose panagrinėsime šį uždavinį detaliau, pradedami paprasčiausio atvejo nagrinėjimu.

12. ĮVERTINYS IR ĮVERTIS. EMPIRINIO VIDURKIO SKIRSTINYS. GRAŽINTINIS ĖMIMAS

1. Pagal Lietuvos standartą **įvertinys** – tai *statistika, kuria naudojamosi populiacijos parametrai vertinti*, o **įvertis** – tai *įvertinio reikšmė, gauta iš konkrečios imties*.

Taigi, *įvertinys* – tai atsitiktinių dydžių funkcija, o *įvertis* – šios funkcijos reikšmė.

Čia ir toliau atsitiktinius dydžius, statistikas, kitas funkcijas žymėsime didžiosiomis raidėmis (pavyzdžiui, $X, \bar{X}, S, Y_i$), o šių atsitiktinių dydžių, statistikų, kitų funkcijų konkrečias reikšmes – mažosiomis raidėmis (pavyzdžiui, $x, \bar{x}, s, y_i$). Taigi, jei **įvertinys**

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

yra imties $X_1, \dots, X_n$ *empirinis vidurkis*, tai **įvertis**

$$\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$$

yra šio vidurkio konkrečioji reikšmė, arba tiesiog reikšmė.

Beje, imties $X_1, \dots, X_n$ konkrečių reikšmių seką (arba tiesiog *konkrečiąją imtį*) $x_1, \dots, x_n$ 2-jame skyrelyje pavadiname stebinių seka, *stebiniais*.

2. Praeitame skyrelyje minėjome, kaip svarbu žinoti *įvertinio* (t.y. statistikos) *skirstinį*. Minėjome, kad tai sudėtingas uždavinys, kurį išnagrinėsime paprasčiausios statistikos – empirinio vidurkio $\bar{X}$ – atveju, pasirinkdami paprastą populiaciją, specialiai pritaikytą statistikos vadovėliui.

3. Tarkime, kad populiacija sudaryta tik iš keturių elementų. Tai 132, 140, 160, 168 – svoriai gramais keturių kriaušių, nupirktų maisto prekių parduotuvėje Vilniuje. Iš šios populiacijos nagrinėjama imtis, kurios dydis yra 2.

Tarkime, kad kriaušės yra sudėtos į dėžę. Tikimybė p ištraukti svorio x kriaušę yra $1/4$. Jei X yra diskretusis atsitiktinis dydis, nusakantis atsitiktinai ištrauktos kriaušės svorį x , tai jo skirstinys pilnai aprašomas tokia lentele:

Svoris x	132	140	160	168
Tikimybė p	1/4	1/4	1/4	1/4

Nesunku apskaičiuoti šio diskretaus skirstinio vidurkį μ ir dispersiją σ^2 :

$$\mu = 132 \cdot \frac{1}{4} + 140 \cdot \frac{1}{4} + 160 \cdot \frac{1}{4} + 168 \cdot \frac{1}{4} = 150,$$

$$\sigma^2 = (132 - 150)^2 \cdot \frac{1}{4} + (140 - 150)^2 \cdot \frac{1}{4} +$$

$$+ (160 - 150)^2 \cdot \frac{1}{4} + (168 - 150)^2 \cdot \frac{1}{4} = 212.$$

4. Toliau yra naudingos kelios naujos sąvokos, kurias suformuluosime pagal Lietuvos standartą.

Ėmimas, rinkimas – tai imties ėmimo ar sudarymo procesas.

Gražintinis ėmimas¹ – tai ėmimas, kai kiekvienas paimtas ir užregistruotas ėmimo vienetas gražinamas į populiaciją prieš imant kitą ėmimo vienetą.

Negražintinis ėmimas² – tai ėmimas, kai ėmimo vienetai iš populiacijos imami tik vieną kartą ar vienas po kito negražinant jų į populiaciją.

5. Pункte 3 apskaičiavome populiacijos parametrus, kai populiacija yra pilnai žinoma. Dabar nagrinėsime gražintinio ėmimo schemą, kai išvados apie populiacijos parametrus daromos turint tik imtį, kurios dydis yra 2.

Jei populiacija sudaryta iš keturių elementų, iš kurių imama imtis didumo 2, tai galimi tik 4×4 , t.y. 16 variantų, kuriuos galima pavaizduoti 8 lentele.

8 lentelė

Konkrečios imties numeris	Konkrečioji imtis ($n = 2$)	Konkretus empirinis vidurkis $\bar{x}$	Tikimybė gražintinis ėmimas
1	132; 132	132	1/16
2	132; 140	136	1/16
3	132; 160	146	1/16
4	132; 168	150	1/16
5	140; 140	140	1/16
6	140; 132	136	1/16
7	140; 160	150	1/16
8	140; 168	154	1/16
9	160; 160	160	1/16
10	160; 132	146	1/16
11	160; 140	150	1/16
12	160; 168	164	1/16
13	168; 168	168	1/16
14	168; 132	150	1/16
15	168; 140	154	1/16
16	168; 160	164	1/16

¹ Angl. *sampling with replacement*; rus. *выбор с возвращением*.

² Angl. *sampling without replacement*; rus. *выбор без возвращения*.

Atrinkę skirtingas $\bar{x}$ reikšmes ir susumuodami pasikartojančių reikšmių tikimybes, nesunkiai gauname empirinio vidurkio $\bar{X}$ skirstinį, kuris pilnai aprašomas 9 lentele:

9 lentelė

Konkretus empirinis vidurkis $\bar{x}$	132	136	140	146	150	154	160	164	168
Tikimybė	1/16	1/8	1/16	1/8	1/4	1/8	1/16	1/8	1/16

6. Žinant statistikos $\bar{X}$ skirstinį, nesunku apskaičiuoti šio skirstinio vidurkį ir dispersiją, kuriuos mes žymėsime atitinkamai $\mu_{\bar{X}}$ ir $\sigma_{\bar{X}}^2$. Svarbu skirti šiuos dydžius nuo analogiškų charakteristikų μ ir σ^2 , aprašančių pačią populiaciją, kuriuos mes apskaičiavome punkte 3.

Taigi, pagal diskretaus atsitiktinio dydžio vidurkio apskaičiavimo taisyklę,

$$\begin{aligned}\mu_{\bar{X}} = & 132 \cdot \frac{1}{16} + 136 \cdot \frac{1}{8} + 140 \cdot \frac{1}{16} + 146 \cdot \frac{1}{8} + 150 \cdot \frac{1}{4} + \\ & + 154 \cdot \frac{1}{8} + 160 \cdot \frac{1}{16} + 164 \cdot \frac{1}{8} + 168 \cdot \frac{1}{16} = 150.\end{aligned}$$

Iš čia darome išvadą, kad $\mu_{\bar{X}} = \mu = 150$.

Analogiškai,

$$\begin{aligned}\sigma_{\bar{X}}^2 = & (132 - 150)^2 \cdot \frac{1}{16} + (136 - 150)^2 \cdot \frac{1}{8} + (140 - 150)^2 \cdot \frac{1}{16} + \\ & + (146 - 150)^2 \cdot \frac{1}{8} + (150 - 150)^2 \cdot \frac{1}{4} + (154 - 150)^2 \cdot \frac{1}{8} + \\ & + (160 - 150)^2 \cdot \frac{1}{16} + (164 - 150)^2 \cdot \frac{1}{8} + (168 - 150)^2 \cdot \frac{1}{16} = 106.\end{aligned}$$

Pastebime, kad šiuo konkrečiu atveju $\sigma_{\bar{X}}^2 = 106$ ir $\sigma^2 = 212$, todėl $\sigma_{\bar{X}}^2 = \sigma^2/2$. Tai teisinga ir bendru atveju:

$$\sigma_{\bar{X}}^2 = \frac{\text{Atsitiktinės imties dispersija}}{\text{Imties dydis } n} = \frac{\sigma^2}{n}.$$

7. Dabar mes „subrendome“ kai kurioms išvadoms.

Jei yra nagrinėjamas gražintinis ėmimas ir imties dydis yra n , o populiacijos vidurkis yra μ , standartinis nuokrypis – σ , tai

$$\mu_{\bar{X}} = \mu = \text{Populiacijos vidurkis},$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{Populiacijos standartinis nuokrypis}}{\sqrt{\text{Imties dydis}}}.$$

Tai reiškia, kad statistika $\bar{X}$ yra pakankamai geras kažkuria prasme populiacijos skirstinio vidurkio μ įvertinys nagrinėjamame tikimybiname modelyje.

Ne mažiau svarbu atkreipti dėmesį į tai, kad konkrečiu atveju detaliai apskaičiuotas statistikos $\bar{X}$ teorinis skirstinys, kuris aprašomas 9 lentelė. Tiesa, išnagrinėtame modelyje populiacija buvo labai paprasta, bet matematikoje yra priimta nuo paprastesnio modelio žengti prie sudėtingesnio.

8. Beje, sąryšis $\mu_{\bar{X}} = \mu$ ir bendru atsitiktinės imties $X_1, \dots, X_n$ atveju yra įrodomas gana paprastai*:

$$\mu_{\bar{X}} = \mathbf{E}\bar{X} = \mathbf{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}X_i = \frac{n\mu}{n} = \mu.$$

Čia pasinaudojome atsitiktinės imties apibrėžimu, pagal kurį atsitiktiniai dydžiai yra vienodai pasiskirstę, ir pažymėjome $\mu = \mathbf{E}X_1 = \dots = \mathbf{E}X_n$.

Formulę $n\sigma_{\bar{X}}^2 = \sigma^2$, gautą konkrečiam uždaviniui, taip pat nesunku įrodyti* ir bendru atveju:

$$\begin{aligned} \sigma_{\bar{X}}^2 &= \mathbf{D}\bar{X} = \mathbf{E}(\bar{X} - \mathbf{E}\bar{X})^2 = \mathbf{E}\left(\frac{X_1 + \dots + X_n}{n} - \mathbf{E}\frac{X_1 + \dots + X_n}{n}\right)^2 = \\ &= \mathbf{E}\left(\frac{X_1 + \dots + X_n}{n} - \mu\right)^2 = \mathbf{E}\left(\frac{(X_1 - \mu) + \dots + (X_n - \mu)}{n}\right)^2 = \\ &= \frac{1}{n^2} \mathbf{E}\left(\sum_{j=1}^n (X_j - \mu)\right)^2 = \frac{1}{n^2} \sum_{j=1}^n \mathbf{E}(X_j - \mu)^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Čia, kaip paprastai, žymime $\mathbf{E}(X_j - \mu)^2 = \sigma^2$. Be to, pasinaudojome sąryšiu

$$\mathbf{E}(X_i - \mu)(X_j - \mu) = 0, \quad \text{jei } i \neq j,$$

kuris yra teisingas bet kuriems nepriklausomiems atsitiktiniams dydžiams X_i, X_j , turintiems vidurkį μ .

* Nestudijavę tikimybių teorijos kurso skaitytojai šį įrodymą gali praleisti.

13. NEGRAŽINTINIS ĖMIMAS. KOREKCIJOS KOEFICIENTAS. VIDURKIO STANDARINĖ PAKLAIDA

1. Grįžkime prie praeitame skyrelyje išnagrinėto pavyzdžio, kai populiacija sudaryta tik iš keturių elementų. Tai 132, 140, 160, 168 – svoriai gramais keturių kriaušių, nupirktų maisto prekių parduotuvėje. Iš šios populiacijos nagrinėjama imtis, kurios dydis 2, tačiau šįkart tai – *negražintinio ėmimo* schema. Tai reiškia, kad iš dėžės atsitiktinai ištraukta kriaušė nėra grąžinama atgal.

Jei nagrinėjamas negražintinis ėmimas, populiacija sudaryta iš keturių elementų, o imties dydis yra 2, tai pirmąjį elementą galima pasirinkti bet kaip, t.y. keturiais būdais, o antrąjį – jau vienu mažiau, t.y. trim būdais. Taigi, galimi tik 4×3 , t.y. 12 variantų, kuriuos galima pavaizduoti 10 lentele.

10 lentelė

Konkrečios imties numeris	Konkrečios imties reikšmės	Konkrečios empirinio vidurkio reikšmė $\bar{x}$	Tikimybė (imtis be grąžinimo)
1	132; 140	136	1/12
2	132; 160	146	1/12
3	132; 168	150	1/12
4	140; 132	136	1/12
5	140; 160	150	1/12
6	140; 168	154	1/12
7	160; 132	146	1/12
8	160; 140	150	1/12
9	160; 168	164	1/12
10	168; 132	150	1/12
11	168; 140	154	1/12
12	168; 160	164	1/12

Atrinkę skirtingas $\bar{x}$ reikšmes ir susumuodami pasikartojančių reikšmių tikimybes, gauname empirinio vidurkio $\bar{X}$ skirstinį, kuris pilnai aprašomas 11 lentele.

2. Šiuo atveju $\bar{X}$ vidurkis $\mu_{\bar{X}}$ yra toks:

$$\mu_{\bar{X}} = 136 \cdot \frac{1}{6} + 146 \cdot \frac{1}{6} + 150 \cdot \frac{1}{3} + 154 \cdot \frac{1}{6} + 164 \cdot \frac{1}{6} = 150.$$

Matome, kad skirstinio $\bar{X}$ vidurkis $\mu_{\bar{X}}$ imties be grąžinimo atveju sutampa su

11 lentelė

Konkreto empirinio vidurkio reikšmė $\bar{x}$	136	146	150	154	164
Tikimybė	1/6	1/6	1/3	1/6	1/6

skirstinio $\bar{X}$ vidurkiu imties su grąžinimu atveju. Ši taisyklė yra teisinga ir bendru atveju:

Ir grąžintinio, ir negrąžintinio ėmimo atvejais empirinio vidurkio $\bar{X}$ matematinės viltys sutampa ir yra lygios teorinio skirstinio matematinei vilčiai μ .

3. Dabar apskaičiuosime dispersiją $\sigma_{\bar{X}}^2$, turėdami galvoje, kad $\mu_{\bar{X}} = 150$:

$$\begin{aligned}\sigma_{\bar{X}}^2 &= (136 - 150)^2 \cdot \frac{1}{6} + (146 - 150)^2 \cdot \frac{1}{6} + (150 - 150)^2 \cdot \frac{1}{3} \\ &\quad + (154 - 150)^2 \cdot \frac{1}{6} + (164 - 150)^2 \cdot \frac{1}{6} = \frac{424}{6}.\end{aligned}$$

Šią formulę galima perrašyti taip:

$$\sigma_{\bar{X}}^2 = \frac{212}{2} \cdot \frac{4 - 2}{4 - 1}.$$

Ši formulė – gera iliustracija bendram atvejui, kai populiacija yra sudaryta iš N elementų, o šios populiacijos pagrindu atliekamas n elementų negrąžintinis ėmimas:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}.$$

Iš čia gauname, kad

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}.$$

Palyginkime šią formulę su dispersija $\sigma_{\bar{X}}^2 = \sigma^2/n$, gauta praeitame skyrelyje grąžintinio ėmimo atveju. Matome, kad jos skiriasi daugikliu

$$\sqrt{(N - n)/(N - 1)},$$

kuris yra vadinamas baigtinės populiacijos korekcijos koeficientu¹.

¹ Angl. *correction factor*, rus. *поправочный коэффициент*.

4. Lengva pastebėti, kad jei populiacija yra labai didelė, lyginant su imties dydžiu n , tai šis koeficientas yra arti vieneto. Pavyzdžiui, jei $N = 4000$, o $n = 10$, tai

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{4,000-10}{4,000-1}} \approx 0,9989.$$

Praktiniuose uždaviniuose šis koeficientas skaitomas lygus 1. Taigi, jei populiacijos elementų skaičius N yra pakankamai didelis imties dydžio n atžvilgiu, tai $\sigma_{\bar{X}} \approx \sigma/\sqrt{n}$ ir todėl galima skaityti, kad $\bar{X}$ standartinis nuokrypis negražintinio ėmimo atveju yra beveik toks pats, kaip ir gražintinio ėmimo atveju.

Intuityviai tai nesunku suvokti. Iš tiesų, jei mes turime ne 4, o 4000 kriausių, ir nagrinėjame imtį, kurios dydis $n = 10$, tai eksperimento rezultatas tuo atveju, kai tas kriaušes negražiname, praktiškai nesiskirs nuo rezultato tuo atveju, kai kriaušes gražiname atgal.

Empirinio vidurkio $\bar{X}$ standartinis nuokrypis paprastai yra vadinamas vidurkio *standartine paklaida*. Taigi, gražintinio ėmimo atveju $\bar{X}$ standartinė paklaida yra $\sigma/\sqrt{n}$, o negražintinio ėmimo atveju $-(\sigma/\sqrt{n})\sqrt{(N-n)/(N-1)}$. Aki-vaizdu, kad juo didesnė imtis, juo mažesnė standartinė paklaida.

5. Pavyzdys. Iš visų mokyklos moksleivių pasirenkami 25 mokiniai. Yra žinoma, kad šių moksleivių ūgių skirstinio standartinis nuokrypis lygus 2. Rasti ūgio vidurkio standartinę paklaidą, jei:

(a) mokykloje yra 1000 moksleivių;

(b) jei tiriami ne vienos mokyklos, o viso pasaulio moksleiviai, t.y. populiacija yra labai didelė.

Sprendimas. (a) Turime, kad $n = 25$, $N = 1000$, $\sigma = 2$. Taigi, vidurkio standartinė paklaida lygi

$$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{2}{\sqrt{25}} \cdot \sqrt{\frac{1000-25}{1000-1}} \approx 0,395.$$

(b) Šiuo atveju vidurkio standartinė paklaida yra lygi

$$\sigma/\sqrt{n} = 2/\sqrt{25} = 0,4.$$

14. EMPIRINĖ CENTRINĖS RIBINĖS TEOREMOS INTERPRETACIJA

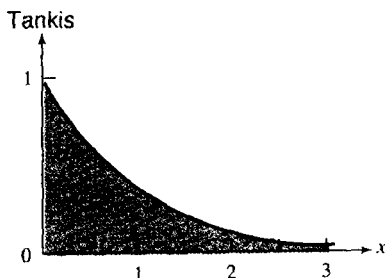
1. 12-me ir 13-me skyreliuose daug dėmesio skyrėme empirinio vidurkio $\bar{X}$ skirstiniui specialiu atveju bei dviems parametrams – vidurkiui $\mu_{\bar{X}}$ ir dispersijai $\sigma_{\bar{X}}$ – apskaičiuoti. Šiame skyrelyje žengsime toliau – nagrinėsime $\bar{X}$ skirstinį bendru atveju.

Kadangi $\bar{X} = (X_1 + \dots + X_n)/n$, tai $\bar{X}$ skirstinys, atrodytų, turi labai priklausyti nuo atsitiktinių dydžių $X_1, \dots, X_n$ skirstinio. Žodžio „atrodytų“ pastarajame sakinyje išvis neturi būti, kai n yra mažas. Mes matėme tai apskaičiuodami $\bar{X}$ skirstinį tiesiogiai, kai populiacija buvo sudaryta iš keturių elementų, iš kurių buvo imama imtis didumo 2.

2. Toks tiesioginis $\bar{X}$ skirstinio radimas bendru atveju, suprantama, nėra paprastas uždavinys. Todėl tikrai nuostabia ir svarbia galima pavadinti teoremą, kuri tvirtina, kad jei n yra pakankamai didelis, tai $\bar{X}$ skirstinys yra artimas normaliajam skirstiniui. Apie tai trumpai buvo kalbama 10-to skyrelio pabaigoje – tai centrinė ribinė teorema. Suformuluosime ją tokioje formoje, kuri yra suprantama kiekvienam, taikančiam statistikos rezultatus praktikoje.

Centrinė ribinė teorema. Tegu $X_1, \dots, X_n$ – praktiškai bet kokia atsitiktinė imtis su vidurkiu μ ir standartiniu nuokrypiu σ . Tada empirinio vidurkio $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ skirstinys gali būti pakankamai gerai aproksimuotas normaliuoju skirstiniu su vidurkiu μ ir standartiniu nuokrypiu $\sigma/\sqrt{n}$, jei tik n yra pakankamai didelis.

3. Norėdami papildyti plačią centrinės ribinės teoremos veikimo sritį, išnagrinėsime pavyzdį, kai ėmimas atliekamas iš populiacijos, kurios skirstinys yra tolygus, nesimetrinis ir kurio tankis $p(x)$ gerokai skiriasi nuo normaliojo:



24 pav. Centrinės ribinės teoremos praktinio „patikrinimo“ pavyzdyje tankis parinktas taip, kad jis gerokai skiriasi nuo normaliojo tankio

Kaip matyti iš brėžinio, pasirinktasis tankis yra eksponentinė funkcija, $p(x) = \exp(-x)$, kai $x > 0$. Tokiu atveju pati populiacija turi taip vadinamą *eksponentinį skirstinį* su vidurkiu $\mu = 1$ ir dispersija $\sigma^2 = 1$. Kompiuteriu galima sumodeliuoti šio skirstinio 200 imčių, kurių kiekvienos dydis $n = 30$. Reikšmių diapazonu pasirenkame 0,1. Gauname 12 numerių pažymėtą $\bar{X}$ skirstinio tankių lentelę (pagal R.Khazanie [6]):

12 lentelė

Reikšmė	Dažnumas	Santykinis dažnis	Santykinio dažnio/diapazono plotis (sant. dažnis /0,1)
0,5	1	0,005	0,05
0,6	3	0,015	0,15
0,7	11	0,055	0,55
0,8	27	0,135	1,35
0,9	38	0,190	1,90
1,0	43	0,215	2,15
1,1	36	0,180	1,80
1,2	20	0,100	1,00
1,3	10	0,050	0,50
1,4	6	0,030	0,30
1,5	2	0,010	0,10
1,6	3	0,015	0,15

Šios tankių lentelės pagrindu konstruojame $\bar{X}$ histogramą. Kadangi $n = 30$, tai tam pačiam grafike palyginimui nubrėžiame normaliojo skirstinio su vidurkiu 1 ir standartiniu nuokrypiu $1/\sqrt{30}$ tankio grafiką (žr. 25 pav.)

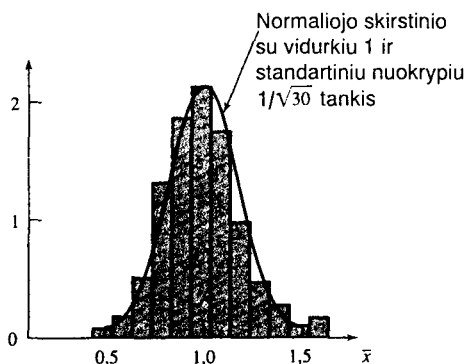
Matome, kad šiuo atveju histograma pakankamai gerai aproksimuojama normaliuoju tankiu.

4. Patogi kita centrinės ribinės teoremos forma, leidžianti ją trumpai suformuluoti taip:

Centrinė ribinė teorema. Jei imties dydis n yra pakankamai didelis, tai bet kuriems realiems a ir b

$$P(a < \bar{X} < b) \approx P\left(\frac{a - \mu}{\sigma/\sqrt{n}} < Z < \frac{b - \mu}{\sigma/\sqrt{n}}\right),$$

t.y. $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ skirstinys yra pakankamai gerai aproksimuojamas standartinio normaliojo atsitiktinio dydžio Z skirstiniu.



25 pav. Centrinės ribinės teoremos „patikrinimo“ rezultatas – vidurkio $\bar{X}$ tankio histograma pakankamai gerai aproksimuojama normaliuoju tankiu

5. Centrinėje ribinėje teoremoje yra prielaida, kad imties didumas n būtų pakankamai didelis. Bet kokią n skaityti pakankamai dideliu?

Atsakymas į šį klausimą priklauso nuo to, kokia papildoma informacija yra žinoma apie populiacijos skirstinį. Ir $n = 1$, ir $n = 2$, ir $n = 3$ gali būti „pakankamai dideli“, jei žinoma, kad populiacijos skirstinys yra normalusis. Suformuluosime tai kaip atskirą teiginį.

Jei populiacijos skirstinys sutampa su normaliuoju skirstiniu, tai bet kuriam natūriniam n imties $X_1, \dots, X_n$ empirinis vidurkis $\bar{X}$ turi normalųjį skirstinį.

Šį teiginį lengva įrodyti* tiems, kurie susipažinę su charakteristinėmis funkcijomis. Iš tiesų, yra duota, kad populiacijos skirstinys, taigi, ir charakteristinė funkcija, yra normalūs:

$$f(t) = \mathbf{E} \exp(itX_j) = \exp(i\mu t - \sigma^2 t^2 / 2).$$

Bet tokiu atveju

$$\begin{aligned} f_{\bar{X}}(t) &= \mathbf{E} \exp(it\bar{X}) = \mathbf{E} \exp\left(\frac{it}{n} \sum_{j=1}^n X_j\right) = \prod_{j=1}^n \mathbf{E} \exp\left(\frac{it}{n} X_j\right) = \\ &= f^n(t/n) = \left(\exp\left(\frac{i\mu t}{n} - \frac{\sigma^2 t^2}{n^2 \cdot 2}\right) \right)^n = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2n}\right), \end{aligned}$$

t.y. $\bar{X}$ turi normalųjį skirstinį su vidurkiu μ ir standartiniu nuokrypiu $\sigma/\sqrt{n}$.

* Nestudijavę tikimybių teorijos kurso skaitytojai šį įrodymą gali praleisti.

Teisingas teiginys, atvirkščias suformuluotajam. Tai seka iš taip vadinamos H. Kramero teoremos apie normaliojo atsitiktinio dydžio dėmenų normalumą.

6. Grįžkime prie suformuluoto klausimo – kokią n centrinėje ribinėje teoroje galima skaityti „pakankamai dideliu“?

Jei populiacijos skirstinys yra simetrinis, tai aproksimacija normaliuoju skirstiniu daugeliu atvejų pakankamai tiksli jau kai $n = 10$.

Bendru atveju aproksimacija normaliuoju skirstiniu yra pakankamai gera, imties dydžiui n pasiekus skaičių 30. Žinoma, juo imties dydis n yra didesnis, juo tikslesnė yra aproksimacija normaliuoju skirstiniu.

7. **Pavyzdys 1.** Firmos gaminamų lempučių veikimo trukmė, skaičiuojama valandomis, yra atsitiktinis dydis, kurio vidurkis yra 500, o standartinis nuokrypis yra 35. Kokia tikimybė, kad atsitiktinė imtis, sudaryta iš 49 elementų, turės vidurkį tarp 488 ir 505?

Sprendimas. Turime, kad $\mu = 500$, $\sigma = 35$, $n = 49$. Taigi, $\mu_{\bar{X}} = 500$ ir

$$\sigma_{\bar{X}} = 35/\sqrt{49} = 5.$$

Mus domina tikimybė $P(488 < \bar{X} < 505)$:

$$\begin{aligned} P(488 < \bar{X} < 505) &\approx P\left(\frac{488 - 500}{5} < Z < \frac{505 - 500}{5}\right) \\ &= P(-2,4 < Z < 1) = 0,3413 + 0,4918 = 0,8331. \end{aligned}$$

Čia mes pasinaudojome standartinio normaliojo skirstinio reikšmių lentelėmis ir toliau trumpai pakomentuosime gautą rezultatą.

Taigi tarkime, kad mes atsitiktinai renkamės lemputes iš elektros lempučių aibės, kaskart pasirinkdami 49 lempučių rinkinį. Apskaičiavę šių lempučių veikimo trukmę, gauname konkrečiąją imtį, kurios dydis $n = 49$. Kiekvienai tokiai imčiai apskaičiuojame empirinį vidurkį. Tada apytiksliai 83% iš nagrinėjamų empirinių vidurkių turės reikšmes tarp 488 ir 505 valandų.

8. **Pavyzdys 2.** Maisto paketo, atsitiktinai paimto iš konteinerio, svoris yra atsitiktinis dydis, kurio vidurkis lygus 452,8 g, o standartinis nuokrypis yra 16,98. Konteineryje yra 36 paketai. Apskaičiuoti tikimybę, kad atsitiktinai pasirinktas konteineris svers daugiau 16,555 kg.

Sprendimas. Turime, kad $\mu = 452,8$; $\sigma = 16,98$ ir $n = 36$. Norime apskaičiuoti tikimybę, kad 36-ių maisto paketų svoris viršys 16555,5 g, t.y.

kad svorio vidurkis $\bar{X}$ viršys $16555,5 \text{ g}/36 = 459,875 \text{ g}$. Pagal centrinę ribinę teoremą ir standartinio normaliojo skirstinio reikšmių lenteles gauname, kad

$$P(\bar{X} > 459,875) \approx P\left(Z > \frac{459,875 - 452,8}{16,98/6}\right) = P(Z > 2,5) = 0,0062.$$

Taigi, ieškomoji tikimybė labai maža, maždaug 62 iš 10 000.

15. „GERIAUSIO ĮVERTINIO“ PROBLEMA. TAŠKINIS IR INTERVALINIS VERTINIMAI. NEPASLINKTASIS ĮVERTINYS

1. Konkrečiuose pavyzdžiuose matėme, kad ištirti visą populiaciją, t.y. kiekvieną jos elementą, gali būti paprasčiausiai pernelyg brangu arba dėl eilės priežasčių tiesiog neįmanoma. Tokiais atvejais yra tik vienas kelias – imties pagrindu pateikti statistines išvadas apie nagrinėjamos populiacijos parametrus, kurios „pakankamai gerai“ atspindi visą populiaciją. Statistinės išvados ir populiacijos parametro sąvokos buvo išnagrinėtos 11-ajame skyrelyje.

2. Minėtos išvados apie parametrus teikiamos, apskaičiavus tam tikrų funkcijų (kurias 11-ajame skyrelyje pavadino *statistikomis*) nuo konkrečios imties reikšmės. Ir nors iki šiol mes faktiškai nagrinėjome vienintelę statistiką – empirinį vidurkį – egzistuoja, suprantama, be galo daug tokių konkrečios imties funkcijų (statistikų), kurias galima pasiūlyti minėtų parametrų įvertiniais. Pakartosime, kad *įvertinys* – tai statistika, kuria naudojamosi populiacijos parametrai vertinti, o *įvertis* – tai įvertinio reikšmė, gauta iš konkrečios imties.

Natūraliai iškyla toks klausimas: *kokia imties funkcija geriausiai įvertina populiacijos parametrai (parametrus)?* Čia nedelsiant iškyla dar vienas klausimas: *ką reiškia sąvoka „geriausias įvertinys“?*

3. Kyla pagunda pavadinti geriausiu tokį įvertinį, kuris kuria nors prasme yra artimiausiai prie vertinamojo populiacijos parametro. Tačiau bet koks įvertinys – tai konkrečių imčių funkcija. Tai reiškia, kad apie įvertinio kokybę reikia spręsti ne pagal atskiras konkrečias jo reikšmes, t.y. ne pagal įverčius, o pagal jo reikšmių didelėje eksperimentų serijoje skirstinį. Jei pagrindinė šio skirstinio masė sukoncentruota mažoje vertinamojo parametro aplinkoje, tai su didele tikimybe galima teigti, kad įvertis mažai skiriasi nuo ieškomojo parametro tikrosios reikšmės.

Šiuo požiūriu įvertinys bus tuo „geresnis“, kuo mažesnė yra sklaida apie tikrąją populiacijos parametro reikšmę.

4. Gyvenime su įverčiu (įvertiniu, vertinimu) turėjome reikalų kiekvienas. Pavyzdžiui, susiruošę keliauti iš Vilniaus į Kauną automobiliu, atstumą tarp šių miestų įvertiname 102 kilometrais, automobilio benzino sąnaudas šiam atstumui įveikti įvertiname 6-iais litrais, 1 litro benzino kainą vertiname 2,10 lito, o iš čia nesunku įvertinti kelionės iš Vilniaus į Kauną ir atgal išlaidas, neskaitant automobilio amortizacijos išlaidų. Toks vertinimo metodas vadinamas *taškiniu vertinimu*.

5. Tačiau galimas kitoks vertinimo būdas, kuris turi kitokių pranašumų. Atstumą tarp Vilniaus ir Kauno įvertiname ne pagal „Lietuvos statistikos metraščio“ duomenis, o apytiksliai, atsižvelgdami į tai, iš kokio Vilniaus rajono ir į kokį rajoną Kaune vykstame. Taigi, mes galime įvertinti šį atstumą tokiu būdu – jis yra intervale tarp 95 km ir 120 km, benzino sąnaudas vertiname tarp 5,5 ir 7,0 litrų, jo kainą – tarp 11 ir 15 Lt. Toks vertinimo metodas vadinamas *intervaliniu vertinimu*.

6. Išnagrinėkime dar vieną pavyzdį. Tarkime, kad mus domina vidutinis mėnesinis priskaičiuotas darbo užmokestis Lietuvos ūkyje 1999 m. balandžio mėnesį. Pažymėkime jį raide p . Suprantama, nėra jokios galimybės kas mėnesį apklausti visus Lietuvos gyventojus, norint rasti tikrąją populiacijos parametro p reikšmę, todėl apsiribojama tam tikro dydžio n konkrečiąja imtimi.

1999 m. gegužės 27 d. „Lietuvos rytas“ skelbia ELTOS duomenis, iš kurių matome, jog naudojant tam tikrą metodiką, nustatyta, kad minėtas darbo užmokestis yra 1068,9 lito. Tai, tačiaugi, dar nereiškia, kad $p = 1068,9$. Tačiau jei būtų žinoma, kad *imties paklaida*¹ yra, pavyzdžiui, $\pm 3,5\%$ (arba $\pm 0,035$), tai galėtume tvirtinti, kad

$$1031,49 \leq p \leq 1106,31.$$

Akivaizdu, kad tam tikra metodika gauta p įverčio reikšmė 1068,9 Lt yra taškinio vertinimo rezultatas, o p priklausomybės intervalui $[1031,49; 1106,31]$ nustatymas – intervalinio vertinimo rezultatas.

7. Pasirenkant populiacijos vidurkio μ įvertinį, dauguma iš mūsų intuityviai suteikia pirmenybę empiriniam vidurkiui $\bar{X}$. Kodėl? Šis intuityvus sprendimas turėtų būti grindžiamas $\bar{X}$ savybėmis, kurias mes išnagrinėjome 12-jame skyrelyje:

1) $\bar{X}$ vidurkis yra μ ;

¹ Angl. *sampling error*, rus. *погрешность выборки*.

2) $\bar{X}$ standartinis nuokrypis yra $\sigma/\sqrt{n}$.

Pirmoji savybė teigia, kad $\bar{X}$ yra kažkuria prasme sucentruotas apie parametą μ , jis nėra niekur pasislinkęs. Tokie įvertiniai vadinami *nepaslinktaisiais įvertiniais*, o bendru atveju šis apibrėžimas formuluojamas taip:

Parametro θ įvertinys θ^* vadinamas *nepaslinktuju įvertiniu*¹, jei įvertinio θ^* vidurkis (matematinė viltis) sutampa su θ .

Iš antrosios savybės matome, kad juo didesnis yra imties dydis n , juo mažesnis yra $\bar{X}$ standartinis nuokrypis, nes daugiklis $\sqrt{n}$ yra vardiklyje. Tai, savo ruožtu, reiškia, kad juo didesnis imties dydis n , juo glaudžiau įvertinys $\bar{X}$ yra sukoncentruotas centro μ aplinkoje, taigi, tuo didesnė tikimybė, kad $\bar{X}$ yra artimas μ .

Taigi, mūsų išvada praktiniams skaičiavimams yra tokia. Jei *konkrečioji imtis* įgyja skaitines reikšmes $x_1, \dots, x_n$ ir jų pagrindu yra gaunama empirinio vidurkio reikšmė $\bar{x} = (x_1 + \dots + x_n)/n$, tai mes sakome, kad $\bar{x}$ yra parametro μ *taškinis nepaslinktasis įvertis*².

16. VIDURKIO PASIKLIOVIMO INTERVALAS, KAI DISPERSIJA YRA ŽINOMA. PAKLAIDOS RĖŽIS

1. Taškinis vertinimas turi rimtų trūkumų. Pavyzdžiui, nėra būdų nustatyti, ar taškinis įvertis yra kažkuria prasme pakankamai geras, jei nėra duota jokios papildomos informacijos.

Alternatyvus metodas taškiniam vertinimui, kaip žinome, yra intervalinis vertinimas, kitaip dar vadinamas pasiklivimo intervalo³ metodu.

Pagal Lietuvos standartą, kai T_1 ir T_2 yra dvi statistikos, θ – vertinamasis populiacijos parametras, o tikimybė

$$P(T_1 \leq \theta \leq T_2)$$

ne mažesnė už $1 - \alpha$ (čia $1 - \alpha$ yra fiksuotas teigiamas skaičius, mažesnis už 1), – tai intervalas tarp T_1 ir T_2 yra parametro θ dvipusis $(1 - \alpha)$ -pasiklivimo intervalas.

¹ Angl. *unbiased estimator*, rus. *несмещенная оценка*.

² Atkreipiame dėmesį, kad $\bar{X}$ yra įvertinys (tai statistika), o $\bar{x}$ – įvertis (tai statistikos reikšmė).

³ Angl. *confidence interval*, rus. *доверительный интервал*.

Kai T yra statistika, θ – vertinamasis populiacijos parametras, o tikimybė $P(T \geq \theta)$ (arba tikimybė $P(T \leq \theta)$) yra ne mažesnė už $1 - \alpha$ (čia $1 - \alpha$ yra fiksuotas teigiamas skaičius, mažesnis už 1), – tai intervalas nuo mažiausios galimos θ reikšmės iki T (arba intervalas nuo T iki didžiausios galimos θ reikšmės) yra parametro θ **vienpusis** $(1 - \alpha)$ -**pasiklivimo intervalas**.

Pasiklivimo intervalo galas T yra statistika, ir todėl apskritai kalbant įgyja skirtingas reikšmes skirtingose imtyse.

Ilgose imčių sekose santykinis dažnis tų atvejų, kai pasiklivimo intervalas aprėpia tikrąją populiacijos parametro reikšmę, yra ne mažesnis už $1 - \alpha$.

Pasiklivimo lygmuo – tai tikimybės reikšmė $(1 - \alpha)$, susijusi su pasiklivimo intervalu.

Pasiklivimo lygmenį dažnai yra patogų išreikšti procentais, t.y. $(1 - \alpha)$ padauginti iš 100 procentų.

2. Detaliau panagrinėsime nežinomo populiacijos vidurkio μ pasiklivimo intervalą, jei yra žinomas šios populiacijos standartinis nuokrypis σ (o tuo pačiu, suprantama, yra žinoma ir populiacijos dispersija).

Prielaida, kad σ yra žinoma, gali atrodyti dirbtinė, nes realiame gyvenime σ dažniausiai nėra žinomas. Šią situaciją panagrinėsime sekančiame skyrelyje, o dabartinis nagrinėjimas yra labai naudingas tolimesniam darbui ir analizei.

3. Nagrinėjant μ pasiklivimo intervalą, priimsime šias prielaidas:

1) Imtis sudaryta iš n nepriklausomų stebėjimų, t.y. eksperimento rezultatai neturi vienas kitam įtakos;

2) Nagrinėjama populiacija turi normalųjį skirstinį su vidurkiu μ , kuris, kaip jau buvo minėta, yra nežinomas;

3) Populiacijos standartinis nuokrypis σ yra žinomas.

Mūsų pirmoji užduotis – sukonstruoti 95 procentų lygmens pasiklivimo intervalą¹, t.y. rasti tokį atsitiktinį intervalą, kad tikimybė, jog μ priklauso šiam intervalui, būtų ne mažesnė už 0,95.

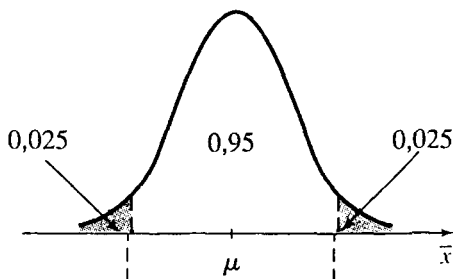
4. 14-me skyrelyje mes matėme, kad jei nagrinėjama populiacija turi normalųjį skirstinį su vidurkiu μ ir dispersija σ^2 , tai empirinio vidurkio $\bar{X}$ skirstinys taip pat yra normalus su vidurkiu μ ir standartiniu nuokrypiu $\sigma/\sqrt{n}$. Standartizuodami gauname, kad statistika

$$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$$

¹ Paraidžiui prisilaikant Lietuvos standarto, reiktų sakyti taip: „sukonstruoti dvipusį 0,95-pasiklivimo intervalą“.

turi standartinį normalųjį skirstinį.

Tam, kad gautume 95 procentų lygmens pasiklovimo intervalą, mes turime iš pradžių rasti tokį z , kad plotas, ribojamas standartinio normalaus tankio kreive zonoje tarp $-z$ ir z , būtų lygus 0,95. Grafiškai tai būtų galima pavaizduoti taip:



26 pav. Kiekviena standartinio normaliojo tankio uodega (kairėje ir dešinėje) lygi 0,025

Iš standartinio normaliojo skirstinio lentelių randame, kad $z = 1,96$. Užrašome gautą rezultatą:

$$P\left(-1,96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1,96\right) = 0,95.$$

Tai reiškia, kad

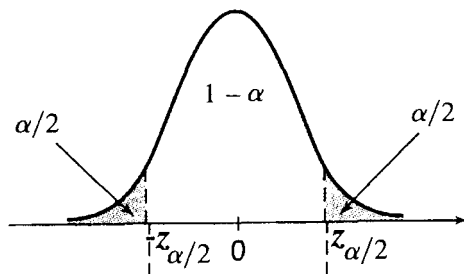
$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95,$$

t.y. *atsitiktinis intervalas* $\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right)$ „dengia“ parametrą μ su tikimybe 0,95.

5. Sukonstravus 95 procentų lygmens pasiklovimo intervalą, nesunku rasti ir $(1 - \alpha) \cdot 100$ procentų lygmens pasiklovimo intervalą. Iš tiesų, 95% lygmens pasiklovimo intervalo atveju $\alpha = 0,05$, o z reikšmę mes radome iš standartinio normaliojo tankio lentelės „uodegos“ reikšmei $0,05/2=0,025$, t.y. $z = z_{0,05/2}$. Analogiškai situacija yra ir bendrojo atveju – plotas tarp $-z_{\alpha/2}$ ir $z_{\alpha/2}$ yra $(1 - \alpha)$. Grafiškai tai galima pavaizduoti taip, kaip atlikta 27 pav.

Taigi, $z_{\alpha/2}$ parinkome taip, kad $P(Z \geq z_{\alpha/2})$, todėl

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$



27 pav. Taškas $-z_{\alpha/2}$ yra standartinio normaliojo atsitiktinio dydžio Z $\alpha/2$ kvantilis, kadangi $P(Z < -z_{\alpha/2}) = \alpha/2 = P(Z \leq -z_{\alpha/2})$

Iš čia darome išvadą, kad su tikimybe $(1 - \alpha)$, t.y. su didele tikimybe įvertis $\bar{X}$ gali skirtis nuo nežinomo parametro μ ne daugiau kaip $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Be to, šis skirtumas gali būti kiek norima mažas, esant pakankamai dideliams imties dydžiui n .

6. Dydis $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ yra vadinamas parametro μ įvertinio maksimalia paklaida $(1 - \alpha) \cdot 100$ procentų lygmenyje arba paklaidos režiu¹.

Gautus rezultatus reziumuosime taip:

Jei populiacija turi normalųjį skirstinį su vidurkiu μ ir dispersija σ^2 ir σ yra žinomas, o μ yra nežinomas, tai parametro μ $(1 - \alpha)$ 100 procentų lygmens pasiklovimo intervalas yra

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

kur $\bar{X}$ yra empirinis vidurkis, o n – imties dydis.

7. Šio skyrelio pradžioje padarėme prielaidą, kad nagrinėjama populiacija turi normalųjį skirstinį. Kam reikėjo šios prielaidos? Tam, kad $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ turėtų standartinį normalųjį skirstinį ir mes galėtume naudotis lentelėmis.

Tačiau jei n yra didelis (30 ir daugiau), tai ši prielaida apie populiacijos normalumą nėra būtina. Iš tiesų, pagal centrinę ribinę teoremą (žr. 14-ąjį skyrelį), $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ skirstinys yra pakankamai gerai aproksimuojamas standartinio normaliojo atsitiktinio dydžio skirstiniu. Tai reiškia, kad $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ turi asimptotiškai normalųjį skirstinį ir mes vėl galime nesunkiai rasti paklaidos režį, tačiau šįkart tai bus *apytikslis* (arba *asimptotinis*) paklaidos režis.

¹ Angl. *margin of error*; rus. *граница погрешности*.

8. Išnagrinėsime konkretų pavyzdį ir pateiksime jo empirinę interpretaciją. Mus domina vidutinis tam tikros profesijos trisdešimtmečių vyrų svoris. Jo reikšmė μ yra nežinoma ir mūsų užduotis yra sukonstruoti šiam parametrai 95% lygmens pasiklovimo intervalą. Iš ankstesnių analogiško pobūdžio eksperimentų buvo padaryta prielaida, kad $\sigma = 6$. Imties dydis n lygus 100. Įstatę žinomas σ ir n reikšmes, randame 95% lygmens pasiklovimo intervalą:

$$\left(\bar{X} - 1,96 \frac{6}{\sqrt{100}}; \bar{X} + 1,96 \frac{6}{\sqrt{100}} \right),$$

t.y. $(\bar{X} - 1,176; \bar{X} + 1,176)$. Taigi,

$$P(\bar{X} - 1,176 < \mu < \bar{X} + 1,176) = 0,95.$$

Tarkime, kad, pasvėrę šimtą trisdešimtmečių vyrų, apskaičiavome jų svorio vidurkį, kuris lygus 79,21 kg. Šiuo atveju

$$(\bar{x} - 1,176; \bar{x} + 1,176) = (79,21 - 1,176; 79,21 + 1,176) = (78,034; 80,386).$$

Pastebėsime, kad būtų klaidinga rašyti

$$P(78,034 < \mu < 80,386) = 0,95$$

bent jau todėl, kad μ nėra atsitiktinis dydis. Pasiklovimo intervalas yra atsitiktinis intervalas.

Ar galima tvirtinti, kad μ priklauso intervalui (78,034; 80,386)? Mes šito nežinome!

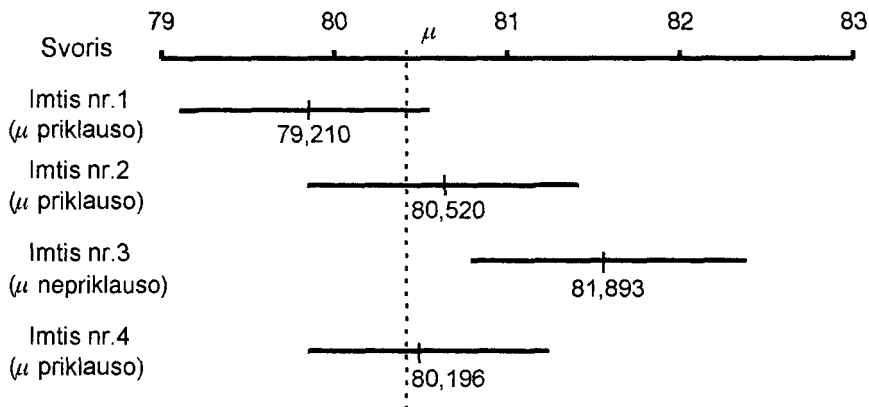
Toliau, tarkime, kad pasvėrę kitus šimtą vyrų, t.y. paėmę kitą imtį, radome, kad jos vidurkis yra 80,52 kg. Ieškomasis intervalas šiuo atveju yra $(80,52 - 1,176; 80,52 + 1,176) = (79,344; 81,696)$. Ir vėl mes negalime tvirtinti, kad μ priklauso būtent šiam intervalui.

Tarkime, kad, tęsdami šį procesą, mes gavome rezultatą, pavaizduotą 13 lentelėje.

13 lentelė

Imties numeris	Konkreiti imties vidurkio reikšmė	Pasiklovimo intervalų reikšmės ($\bar{x} - 1,176; \bar{x} + 1,176$)
1	79,210	(78,034; 80,386)
2	80,520	(79,344; 81,696)
3	81,893	(80,717; 83,069)
4	80,196	(79,020; 81,372)
⋮	⋮	⋮

Kuriems iš šių intervalų priklauso μ , o kuriems – ne, mes negalime tiksliai atsakyti. Tačiau mes galime tvirtinti, kad apytiksliai 95% tokiu būdu sukonstruotiems intervalams priklauso mūsų nagrinėjamos populiacijos vidutinis svoris μ . Grafiškai tai būtų galima pavaizduoti taip:



28 pav. Pasiklovimo intervalų reikšmės. Jei μ būtų tokia, kaip pavaizduota, tai galėtume tvirtinti, kad ji nepriklauso tik trečiajam intervalui

Žinoma, imtis didumo n kartojama tik teorijoje. Konstruojant pasiklovimo intervalą praktikoje, šis procesas nekartojamas. Imama tik viena imtis ir apskaičiuojami abu intervalo galiniai taškai. Populiacijos parametras μ arba priklauso šiam intervalui, arba nepriklauso. O kokia yra tikroji μ reikšmė, kaip sako statistikos profesionalai, praktikoje niekada nesužinosime.

17. STUDENTO SKIRSTINYS. VIDURKIO PASIKLOVIMO INTERVALAS, KAI DISPERSIJA YRA NEŽINOMA

1. Praeitame skyrelyje nagrinėjome nežinomo populiacijos vidurkio μ pasiklovimo intervalą su prielaida, kad šios populiacijos standartinis nuokrypis σ yra žinomas. Kaip jau buvo minėta, praktikoje dažniau sutinkama situacija,

kai σ nėra žinomas. Šį atvejį toliau detaliau išnagrinėsime vėl tardami, kad išpildomos šios dvi prielaidos:

- 1) Imtis sudaryta iš nepriklausomų stebėjimų;
 - 2) Populiacija turi normalųjį skirstinį.
2. Nagrinėdami pasiklivimo intervalą ir paklaidos režius atveju, kai σ yra žinomas, pradėjome nuo statistikos

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

kuri turi standartinį normalųjį skirstinį.

Kadangi dabar σ yra nežinomas, tai logiška būtų sekantį žingsnį žengti, keičiant populiacijos dispersiją σ^2 jos empirine dispersija S^2 (žr. 7-ąjį skyrelį). Pagal Lietuvos standartą priminsime, kad **dispersija** S^2 – tai *sklaidos matas, kuris yra stebinių reikšmių nuokrypių nuo vidurkio kvadratų suma, padalyta iš vienetu sumažinto stebinių skaičiaus*:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Imties dispersija yra populiacijos dispersijos nepaslinktasis įvertis.

Z išraiškoje pakeitę σ į S gauname, kad logiška būtų nagrinėjimą pradėti nuo statistikos

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

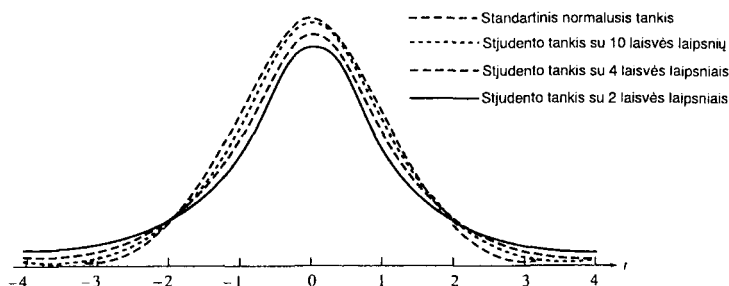
Šios statistikos skirstinys vadinamas *Stjudento skirstiniu* (arba, sutrumpintai, *t skirstiniu*) su $n - 1$ *laisvės laipsniu*¹, kur n yra imties dydis.

3. Stjudento skirstinys turi šias savybes:

- I. Jis yra tolydus, o jo tankis $p(x)$ yra simetrinis ordinačių ašies atžvilgiu.
- II. Jo tankis $p(x)$ turi varpo formą, primenančią normalųjį tankį;
- III. Didėjant laisvės laipsniams, t skirstinys tampa vis labiau panašus į standartinį normalųjį skirstinį.

¹ Angl. *Student's t distribution with $n - 1$ degrees of freedom*.

Visas šias išvadas gerai iliustruoja šis grafikas:

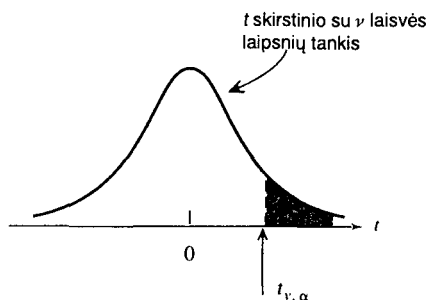


29 pav. Didėjant laisvės laipsniams, Stjudento tankis tampa vis labiau panašus į normalųjį tankį

4. Tegu T yra atsitiktinis dydis, turintis Stjudento skirstinį su ν laisvės laipsnių. Dydis $t_{\nu, \alpha}$ vadinamas Stjudento skirstinio lygmens α kritine reikšme, jei

$$P(T \geq t_{\nu, \alpha}) = \alpha.$$

Sritį, kurios plotas yra α , grafiškai galima būtų pavaizduoti taip (piešinyje ši sritis yra užtušuota):



30 pav. $t_{\nu, \alpha}$ yra lygmens α kritinė reikšmė

Kritinės Stjudento skirstinio reikšmės atvejams $\alpha = 0,01$; $\alpha = 0,05$; $\alpha = 0,025$; $\alpha = 0,01$; $\alpha = 0,005$ bei $\nu = 1, 2, 3, \dots, 30$ yra pateikiamos 4 lentelėje vadovėlio gale.

Iš šios lentelės, beje, galima padaryti tokią išvadą:

Jei Stjudento skirstinio laisvės laipsnių skaičius ν yra lygus 30 arba daugiau, tai vietoj jo apytiksliai galima naudoti standartinį normalųjį skirstinį.

5. Grįžkime prie nežinomo populiacijos vidurkio μ pasiklivimo intervalo konstravimo, kai σ yra nežinomas. Kadangi pagal apibrėžimą T yra Stjudento

skirstinys su $n - 1$ laisvės laipsniu, tai

$$P\left(-t_{n-1, \alpha/2} < T = \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1, \alpha/2}\right) = 1 - \alpha.$$

Iš čia akivaizdu, kad

$$P\left(\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Gautus rezultatus režiuruosime taip:

Jei populiacija turi normalųjį skirstinį su vidurkiu μ ir dispersija σ^2 , kurie abu yra nežinomi, tai vidurkio μ $(1 - \alpha) \cdot 100$ procentų lygmens pasiklovimo intervalas yra

$$\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}},$$

kur n – imties dydis, o $\bar{X}$ ir S yra atitinkamai empirinis vidurkis ir empirinis standartinis nuokrypis.

Jei, be to, $n \geq 30$, tai $t_{n-1, \alpha/2}$ apytiksliai galima pakeisti dydžiu $z_{\alpha/2}$, t.y. apytiksliai

$$\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}},$$

kur skaičius $z_{\alpha/2}$ randamas iš standartinio normaliojo atsitiktinio dydžio Z tankio reikšmių lentelės, $P(Z \geq z_{\alpha/2}) = \alpha/2$.

6. Pavyzdys. Ištyrus 25 vienos rūšies cigaretės laboratorijoje buvo nustatyta, kad vidutinis nuodingų dervų skaičius $\bar{x}$ yra lygus 18,6 miligramų ir, be to, $s = 1,81$. Rasti nuodingų dervų šios rūšies cigaretėse vidurkio μ 95 procentų lygmens pasiklovimo intervalą.

Sprendimas. Šiuo atveju $n = 25$, $\bar{x} = 18,6$, $s = 1,81$ ir $\alpha = 0,05$. Naudodamiesi Stjudento skirstinio (su $\nu = 25 - 1 = 24$ laisvės laipsniais) lentele randame, kad $t_{24, 0,025} = 2,064$. Taigi, ieškomasis 95 procentų lygmens pasiklovimo intervalas yra

$$18,6 - 2,064 \cdot \frac{1,81}{\sqrt{25}} < \mu < 18,6 + 2,064 \frac{1,81}{\sqrt{25}}$$

arba, atlikus aritmetinius veiksmus, $17,87 < \mu < 19,35$.

Iš čia darome išvadą, kad ištirtos rūšies cigarečių nuodingų dervų vidurkis μ yra 95 procentų lygmens pasiklovimo intervale $(17,85; 19,35)$.

18. CHI KVADRATO SKIRSTINYS. DISPERSIJOS PASIKLIVIMO INTERVALAS

1. Norėdami sukonstruoti dispersijos pasiklivimo intervalą, priimsime to-
kias pat prielaidas, kaip ir praeituose dviejuose skyreliuose:

- 1) Imtis sudaryta iš nepriklausomų stebėjimų;
- 2) Populiacija turi normalųjį skirstinį su dispersija σ^2 .

2. Tolimesniam tyrimui naudinga tokia statistika:

$$U = \frac{n-1}{\sigma^2} S^2, \quad (1)$$

kur, priminsime, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Kadangi populiacija turi normalųjį skirstinį su dispersija σ^2 , o prieš S^2 yra daugiklis $1/\sigma^2$, tai galima parodyti, kad statistikos U skirstinys nepriklauso nuo σ ir turi gerai žinomą skirstinį, vadinamą *chi kvadrato skirstiniu*¹ su $n-1$ laisvės laipsniais.

Galima parodyti, kad *chi kvadrato skirstinys* – tai *nepriklausomų normalių standartinių atsitiktinių dydžių kvadratų sumos skirstinys*.

3. Chi kvadrato skirstinys dar yra vadinamas Helmerto–Pirsono skirstiniu. Helmerto (F.R.Helmert) darbas šia tema pasirodė 1875 metais, o Pirsono (K.Pearson) – 1900 metais. Suprantama, kad chi kvadrato skirstinio savybės ir jo taikymo praktiniams uždaviniams galimybės šiandien yra plačiai išnagrinėtos. Kaip ir Stjudento skirstinio atveju, pateiksime paprasčiausias ir pagrindines chi kvadrato skirstinio savybes:

I. Jis yra tolydus ir nesimetrinis ordinačių ašies atžvilgiu;

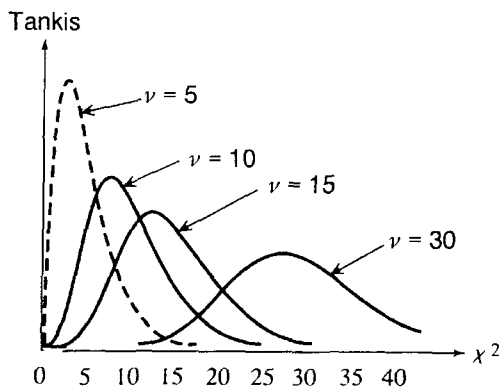
II. Jo tankio forma priklauso nuo parametro ν , kuris yra vadinamas laisvės laipsniais;

III. Didėjant laisvės laipsniams, jis tampa vis labiau panašus į normalųjį skirstinį. Kai $\nu \geq 30$, jis yra apytiksliai normalus.

Visas šias išvadas gerai iliustruoja chi kvadrato skirstinių su ν laisvės laipsniais tankių $p(x)$ grafikas (31 pav.).

4. Tegu χ^2 yra atsitiktinis dydis, turintis chi kvadrato skirstinį su ν laisvės laipsniais. Dydis $\chi^2_{\nu, \alpha}$ vadinamas chi kvadrato skirstinio *lygmens α kritine*

¹ Angl. *Chi-square distribution* (χ^2 distribution); rus. *ху-квадратное распределение* (χ^2 -распределение).

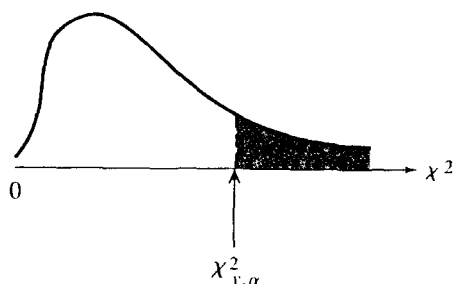


31 pav. Chi kvadrato tankių kreivės

reikšme, jei

$$P(\chi^2 \geq \chi_{\nu, \alpha}^2) = \alpha.$$

χ^2 skirstinio su ν laisvės laipsniais tankio sritis, kurios plotas yra α , piešinyje yra užtūšuota:



32 pav. $\chi_{\nu, \alpha}^2$ yra chi kvadrato skirsnio su ν laisvės laipsnių lygmens α kritinė reikšmė. Užtūšotos srities plotas lygus α

Chi kvadrato skirstinio kritinės reikšmės atvejams $\alpha = 0,995$; $\alpha = 0,99$; $\alpha = 0,975$; $\alpha = 0,95$; $\alpha = 0,05$; $\alpha = 0,025$; $\alpha = 0,01$; $\alpha = 0,005$ bei $\nu = 1, 2, 3, \dots, 30$ yra pateikiamos 3 lentelėje vadovėlio gale.

Analogiškai kaip ir Stjudento skirstinio atveju, iš šios lentelės galima padaryti tokią išvadą:

Jei chi kvadrato skirstinio laisvės laipsnių skaičius ν yra lygus 30 arba daugiau, tai vietoj jo apytiksliai galima naudoti normalųjį skirstinį.

Tačiau chi kvadrato skirstinio tankis, skirtingai nei Stjudento skirstinio tankis, nėra simetrinis, todėl chi kvadrato skirstinio atveju vien reikšmių $\alpha = 0,005$; $\alpha = 0,01$; $\alpha = 0,025$; $\alpha = 0,05$ nepakanka, norint žinoti kritines reikšmes atvejams $1 - \alpha$. Štai kodėl šios reikšmės atvejams 0,995; 0,99; 0,975; 0,95 pateikiamos atskirai.

5. Grįžkime prie dispersijos σ^2 pasiklovimo intervalo konstravimo. Kadangi statistika U , apibrėžiama formule (1), turi chi kvadrato skirstinį su $n - 1$ laisvės laipsniu, tai pagal kritinės reikšmės apibrėžimą

$$P(U \geq \chi_{n-1,1-\alpha/2}^2) = 1 - \alpha/2, \text{ t.y. } P(U < \chi_{n-1,1-\alpha/2}^2) = \alpha/2.$$

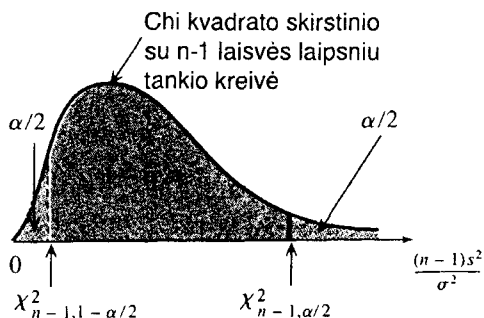
Iš kitos pusės, analogiškai

$$P(U < \chi_{n-1,\alpha/2}^2) = 1 - P(U \geq \chi_{n-1,\alpha/2}^2) = 1 - \alpha/2.$$

Tai reiškia, kad

$$P\left(\chi_{n-1,1-\alpha/2}^2 \leq U = \frac{(n-1)S^2}{\sigma^2} < \chi_{n-1,\alpha/2}^2\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Grafškai tai galima būtų pavaizduoti taip:



33 pav. Chi kvadrato skirstinio kritinių reikšmių radimas

Atlikę elementarias algebrines operacijas gauname, kad

$$P\left(\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2}\right) = 1 - \alpha.$$

Gautą rezultatą reziumuosime taip:

Jei populiacija turi normalųjį skirstinį su nežinoma dispersija σ^2 , tai jos $(1 - \alpha) \cdot 100$ procentų lygmens pasiklivimo intervalas yra

$$\frac{(n-1)S^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha/2}^2},$$

kur S^2 yra imties didumo n empirinė dispersija.

6. Pavyzdys. Mokytojas klasėje atsitiktinai parinko 6 mokinius ir atliko tokį eksperimentą. Mokiniam buvo pasiūlyta atlikti užduotį, o jos atlikimo laikas (minutėmis) buvo fiksuojamas. Gauti tokie rezultatai:

$$8,0 \quad 7,0 \quad 8,5 \quad 7,5 \quad 9,0 \quad 8,0.$$

Apskaičiuoti dydžio, reiškiančio užduoties atlikimo laiką, dispersijos 98% lygmens pasiklivimo intervalą.

Sprendimas. Nesunku įsitikinti, kad $\bar{x} = 8,0$,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^6 (x_i - \bar{x})^2 = \frac{2,5}{5} = 0,5.$$

Kadangi $n = 6$, tai chi kvadrato skirstinys turi 5 laisvės laipsnius. Kadangi prašoma apskaičiuoti 98% lygmens pasiklivimo intervalą, tai $\alpha = 0,02$, o $\alpha/2 = 0,01$; $1 - \alpha/2 = 0,99$. Iš lentelių randame, kad

$$\chi_{5,0,01}^2 = 15,086, \quad \chi_{5,0,99}^2 = 0,554.$$

Tai reiškia, kad

$$\frac{5 \cdot 0,5}{15,086} < \sigma^2 < \frac{5 \cdot 0,5}{0,554}, \quad \text{t.y.} \quad 0,166 < \sigma^2 < 4,513.$$

Iš čia darome išvadą, kad dydžio, reiškiančio užduoties atlikimo laiką, konkrečios imties vidurkis yra 8,0 min., šio dydžio dispersijos σ^2 taškinis įvertis s^2 įgyja reikšmę 0,5, o σ^2 98% lygmens pasiklivimo intervalas yra (0,166; 4,513).

Mes baigėme nagrinėti paprasčiausius statistinių įverčių teorijos klausimus, pavyzdžius ir taikymus praktikoje. Matematiniai šios teorijos aspektai plačiau nagrinėjami antroje šio vadovėlio dalyje.

Sekančiose dviejuose skyreliuose elementariai panagrinėsime hipotezių apie populiacijos vidurkį tikrinimo problemas.

19. HIPOTEZIŲ FORMULAVIMAS. HIPOTEZIŲ TIKRINIMO MECHANIZMO ANALIZĖ

1. Kiekvieną spėjimą apie nežinomą skirstinį arba žinomo skirstinio parametrus vadiname hipoteze arba, tiksliau, *statistine hipoteze*. Tai teiginiai tipo „Nežinomas skirstinys turi savybę D “ arba „Skirstinio nežinomas parametras θ priklauso aibei Ω_0 “ arba pan.

Pastarąjį teiginį panagrinėkime detaliau. Jei Ω yra visų galimų parametro θ reikšmių aibė ir $\Omega = \Omega_0 \cup \Omega_1$ (Ω_0 ir Ω_1 yra nesikertančios aibės, t.y. $\Omega_0 \cap \Omega_1 = \emptyset$), tai galimos tik dvi konkuruojančios hipotezės:

$$H_0: \theta \in \Omega_0 \quad \text{arba} \quad H_1: \theta \in \Omega_1.$$

Hipotezė H_0 vadinama *nuline hipoteze*, o hipotezė H_1 vadinama *alternatyviaja hipoteze*. Jei alternatyviųjų hipotezių yra kelios, naudojamas žymėjimas H_α .

2. Kaip galima būtų paaiškinti termino „nulinė hipotezė“ pasirinkimą? Matomai jis pasirinktas todėl, kad H_0 paprastai žymima hipotezė, kurią tyrinėtojas tikisi atmesti. Žodis „nulinė“ šiame kontekste reiškia, kad teiginys, kurio teisingumą mes siekiame patikrinti, yra niekinis, neteisingas.

Taip termino „nulinė hipotezė“ pasirinkimas grindžiamas vadovėliuose, parašytuose anglų kalba. Vadovėliuose rusų kalba yra prigijęs terminas „pagrindinė hipotezė“, kuris naudojamas kaip *nulinės hipotezės* sinonimas.

3. **Pavyzdys.** Kai kurie vėžiniai augliai nustoja augti be jokio gydymo. Šių auglių kiekio santykį su visų vėžinių auglių kiekiu pažymėkime p_0 . Tegu X yra atsitiktinis dydis, įgyjantis reikšmę 1 arba 0 priklausomai nuo to, ar atsitiktinai pasirinktas auglys, pritaikius naująją terapiją, nustojo augti ar vystėsi toliau. Atsitiktinio dydžio X vidurkis p (taip vadinamasis palankaus įvykio parametras p) yra lyginamas su baziniu parametru p_0 .

Jei $p \leq p_0$ (matyt, tiksliau nagrinėti atvejį $p = p_0$, nes jei $p < p_0$, tai naujoji terapija yra kenksminga), tai terapija yra neefektyvi, o jei $p > p_0$, tai terapija yra efektyvi. Taigi, turime dvi hipotezes:

Nulinė hipotezė $H_0: p \leq p_0$, t.y. terapija yra neefektyvi;

Alternatyvioji hipotezė $H_1 : p > p_0$, t.y. terapija yra efektyvi.

Kaip patikrinti, kuri iš šių dviejų hipotezių yra teisinga, o kuri ne, ir koks galėtų būti hipotezių tikrinimo mechanizmas? Tuo tikslu detaliau panagrinėsime kitą pavyzdį.

4. Gamykla gamina elektros lemputes¹. Jos inžinieriai pasiūlė naują technologiją ir tikisi, kad lempučių darbo trukmė pailgės. Šiuo metu šios trukmės vidurkis yra 450 valandų. Natūrali prielaida apie tai, kad atsitiktinis dydis, išreiškiantis lemputės darbo trukmę iki jos perdegimo, turi normalųjį skirstinį. Be to, yra žinoma, kad standartinis nuokrypis lygus 60 valandų ir šis nuokrypis negali pakisti net po inžinierių išradimo įdiegimo.

Jei inžinierių darbas buvo prasmingas (o tai gana natūrali prielaida), tai lemputės darbo trukmės iki perdegimo vidurkis μ negalėjo sumažėti. Mus domina klausimas, ar po naujos technologijos įvedimo μ padidės, ar liks toks, koks buvo. Taigi, hipotezes galima suformuluoti taip:

Nulinė hipotezė $H_0 : \mu = 450$, t.y. nauja technologija nedavė laukiamų rezultatų;

Alternatyvioji hipotezė $H_1 : \mu > 450$, t.y. nauja technologija pailgino darbo trukmę.

5. Norėdami patikrinti naują inžinierių pasiūlymą, paimkime 9 lemputes, pagamintas naujuoju būdu, ir fiksuokime jų pilną darbo trukmę iki perdegimo. Tarkime, kad gavome tokius rezultatus:

583 518 464 384 445 534 549 563 376.

Išanalizavę šiuos duomenis matome, kad 6 lemputės degė ilgiau už nulinėje hipotezėje fiksuotą skaičių, lygų 450. Iš kitos pusės, 3 lemputės degė trumpiau nei 450 valandų, iš jų dvi degė gerokai trumpiau.

6. Kokį sprendimą galima priimti šių rezultatų pagrindu? Šios imties lempučių darbo trukmės vidurkis yra $\bar{x} = 490,67$. Taigi, jis didesnis už 450. Tai galėtų būti rimtu argumentu naujosios technologijos naudai, jei ši vidurkio reikšmė nėra atsitiktinė svyravimo ar nepastovumo pasekmė.

Kitaip tariant, ar lemputės galėtų būti iš tos lempučių populiacijos, kurioje darbo trukmės vidurkis yra 450 valandų, bet mūsų imties vidurkis atsitiktinai įgijo reikšmę 490,67? Ar ši vidurkio reikšmė yra didesnė už 450 dėl tos priežasties, kad naujoji technologija iš tiesų yra progresyvesnė ir leidžia gaminti ilgaamžiškesnes lemputes?

¹ Dėstoma pagal R.Khazanie [6].

7. Panagrinėsime pirmąjį iš pastarųjų dviejų klausimų. Tuo tikslu apskaičiuosime tikimybę įvykio, jog imties vidurkis įgyja reikšmę 490,67 ar daugiau, jei nulinė hipotezė H_0 yra teisinga, t.y. jei populiacijos vidurkis yra 450 valandų.

Jau minėjome, kad nagrinėjama populiacija turi normalųjį skirstinį su standartiniu nuokrypiu 60. Taigi, jei H_0 yra teisinga, tai empirinio vidurkio $\bar{X}$ skirstinys yra normalus su vidurkiu 450 ir standartiniu nuokrypiu $\sigma/\sqrt{n} = 60/\sqrt{9} = 20$. Taigi,

$$P(\bar{X} \geq 490,67) = P\left(Z \geq \frac{490,67 - 450}{20}\right) = P(Z \geq 2,03) = 0,0212.$$

Taigi, tikimybė įvykio, jog imties vidurkis įgyja reikšmę 490,67 ar daugiau, jei H_0 yra teisinga, yra apytiksliai lygi dviem šimtosioms.

Maža apskaičiuotos tikimybės reikšmė 0,0212 kelia abejones dėl nulinės hipotezės H_0 teisingumo. Bet ar ši tikimybė yra pakankamai maža tam, kad gamyklos vadovai priimtų inžinierių siūlomą naująją technologiją? Vadovų sprendimas priklausys, suprantama, ir nuo tokių praktinių aplinkybių, kaip rekonstrukcijos kaina, jos trukmė, naujųjų lempučių savikaina, jų prognozuojama paklausa ir t.t.

8. Galimas ir kitoks hipotezių tikrinimo metodas. Labai dažnai statistikos specialistai atsižvelgia į rizikos faktorius ir numato juos, pasirinkdami taisyklę išvados (apie hipotezių priėmimą ar atmetimą) priėmimui. Ši taisyklė, kaip pamatysime kitame skyrelyje, vadinama *sprendimo priėmimo taisykle*. Pailiustruosime šį hipotezių tikrinimo metodą aukščiau išnagrinėtu pavyzdžiu apie lempučių darbo trukmę.

Kadangi alternatyvioji hipotezė H_1 teigia, kad $\mu > 450$, tai didelės reikšmės $\bar{x}$ nuteikia mus atmesti H_0 ir priimti H_1 . Juo didesnė reikšmė $\bar{x}$, juo didesnis šis mūsų nusiteikimas. Tarkime, jog mes nusprendėme fiksuoti apatinę vidurkio reikšmę taške 475 ir parinkome tokią sprendimo priėmimo taisyklę:

Atmesti H_0 , jei imties vidurkio reikšmė $\bar{x}$ yra didesnė už 475 valandas.

Mūsų pavyzdyje 9 stebinių imties vidurkis yra 490,67, taigi, jis yra didesnis už 475, ir todėl mes atmetame nulinę hipotezę H_0 ir priimame inžinierių pasiūlymą.

Jei šis vidurkis būtų ne 490,67, o, tarkime, 470 valandų, tai mes nulinę hipotezę H_0 priimtume, o alternatyviąją hipotezę H_1 atmestume.

20. SPRENDIMO PRIĖMIMO TAISYKLĖ. KRITINĖ SRITIS

1. Reziumuosime tai, apie ką buvo kalbama praeitame skyrelyje.

Hipotezių H_0 ir H_1 formulavimas. Jei tikslas yra suformuluoti ir patikrinti teiginį duotosios imties pagrindu, tai šio teiginio neigimas yra nagrinėjamas kaip nulinė hipotezė H_0 , o pats teiginys nagrinėjamas kaip alternatyvioji hipotezė H_1 .

Pavyzdyje apie elektros lempučių naujos gamybos technologijos įdiegimo tikslingumą matėme, kad maža tikimybės $P(\bar{X} > 490, 67)$ reikšmė 0,0212 kelia abejones dėl nulinės hipotezės H_0 teisingumo. Panašiais samprotavimais buvo parengtas kelias sprendimo priėmimo taisyklės išvedimui ir bendru atveju. Apie tai detaliau kalbėsime šiame skyrelyje.

2. **Sprendimo priėmimo taisyklė¹** (arba nulinės hipotezės patikrinimas) – tai procedūra, nusakanti, kokia informacija imties pagrindu ir koku būdu turi būti panaudojama, priimant sprendimą. *Sprendimai gali būti šie:* arba

1° atmesti H_0 ir padaryti išvadą, kad H_1 yra priimta, t.y. priimti H_1 ,
arba

2° priimti H_0 ir padaryti išvadą, kad H_1 yra nepriimtina, t.y. atmesti H_1 .

Svarbu pabrėžti, kad sprendimas atmesti arba priimti H_0 turi būti priimamas griežtos ir akivaizdžios procedūros pagrindu. Pailiustruosime tai konkrečiu pavyzdžiu.

3. **Pavyzdys.** Piko valandomis prie bankomato susidarydavo nemaža klientų eilutė. Norėdamas išvengti šito nepageidaujamo reiškinių, bankas patobulino bankomatą ir nori įsitikinti, kad naujasis modelis dirba efektyviau. Tuo tikslu buvo fiksuojamas klientų aptarnavimo laikas ir nagrinėjama imtis, kurios dydis $n = 38$, bei apskaičiuojamas klientų aptarnavimo laiko vidurkis $\bar{X}$. Yra žinoma, kad bankomato be patobulinimų klientų aptarnavimo laiko skirstinio vidurkis yra 270 sekundžių, o standartinis nuokrypis yra 24. Ar atliktos imties pagrindu galima tvirtinti, kad bankomato su patobulinimais klientų aptarnavimo laiko skirstinio vidurkis μ yra mažesnis už 270 sekundžių?

Sprendimas. Akivaizdu, kad alternatyvioji hipotezė H_1 yra formuluojama taip:

$$H_1: \mu < 270.$$

¹ Angl. *decision rule*, rus. *правило принятия решения*.

Kadangi bankomatas buvo tobulinamas, tarsime, kad atvejis $\mu > 270$ yra negalimas. Tai reiškia, kad nulinė hipotezė H_0 yra tokia:

$$H_0: \mu = 270.$$

Sprendimas čia gali būti toks:

1° atmesti $H_0: \mu = 270$ ir priimti $H_1: \mu < 270$;

2° priimti $H_0: \mu = 270$ ir atmesti $H_1: \mu < 270$.

Sekantis etapas – atsakyti į klausimą, kokioms $\bar{X}$ reikšmėms $\bar{x}$ priimamas sprendimas 1°, t.y. hipotezė H_0 atmetama, ir kokioms $\bar{X}$ reikšmėms $\bar{x}$ priimamas sprendimas 2°, t.y. hipotezė H_0 priimama.

4. Analogiškas klausimas mums jau buvo iškilęs praėjusio skyrelyje. Pavyzdyje apie elektros lemputes matėme, kad maža tikimybė $P(\bar{X} \geq 490,67)$ reikšmė 0,0212 kelia abejones dėl nulinės hipotezės H_0 teisingumo. Ten pat buvo trumpai išnagrinėta sprendimo priėmimo taisyklė, kurioje atsižvelgiama į rizikos faktorius.

Tačiau galimas ir kitoks kelias. Kadangi pavyzdyje apie bankomatus mus domina, ar tikrai $\mu < 270$, t.y. kairioji parametro μ apibrėžimo sritis, tai tik tos $\bar{X}$ reikšmės, kurios yra iš šios kairiosios srities, prieštarauja hipotezei H_0 hipotezės H_1 naudai.

Tarkime, kad mums pavyko rasti tokį skaičių c , kad $P(\bar{X} < c) = \alpha$, kur α – pakankamai mažas skaičius. Tada *sprendimo priėmimo taisyklė* galėtų būti tokia:

atmesti H_0 , jei $\bar{x} < c$;

priimti H_0 , jei $\bar{x} \geq c$.

5. Šią sprendimo priėmimo taisyklę patogiau sutrumpintai žymėti taip:

$$\mathcal{R}: \bar{X} < c,$$

kur $\mathcal{R}$ reiškia hipotezės H_0 atmetimą.

Šiame kontekste baigčių aibė $\{\bar{X} < c\}$ vadinama *atmetimo* (arba *kritine*) *sritimi*¹, o taškas c – *kritine reikšme*.

Dabar panagrinėkime atvejį, kai hipotezė H_0 yra teisinga, t.y. $\mu = 270$. Tada H_0 atmetimas yra neteisingas sprendimas, galintis sąlygoti rimtą klaidą. Kad tokios klaidos nebūtų, mes turime būti tikri, kad tikimybė $P(\bar{X} < c)$ yra labai maža, kai $\mu = 270$. Tarkime, kad mus pilnai tenkina patikimumas $\alpha = 0,05$; t.y. ši tikimybė „labai maža“, jei ji neviršija lygio $\alpha = 0,05$. Tada

¹ Angl. *rejection region*, *critical region*; rus. *область отбраcывания*.

mūsų uždavinys yra rasti tokį c , kad

$$P(\bar{X} < c) = 0,05, \text{ kai } \mu = 270.$$

Mes žinome, kad pakankamai dideliems n $\bar{X}$ turi apytiksliai normalų skirstinį su vidurkiu μ ir standartiniu nuokrypiu $\sigma/\sqrt{n}$. Mūsų atveju $n = 38$, taigi, pakankamai didelis (nes $n > 30$). Be to, mes tarėme, kad $\sigma = 24$. Taigi, kai $\mu = 270$, tai $\bar{X}$ skirstinys yra normalusis skirstinys su vidurkiu 270 ir standartiniu nuokrypiu $24/\sqrt{38}$, o atsitiktinis dydis

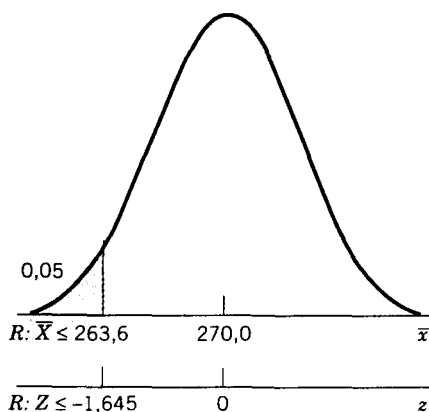
$$Z = \frac{\bar{X} - 270}{24/\sqrt{38}}$$

turi standartinį normalųjį skirstinį.

Iš standartinio normaliojo skirstinio tankio reikšmių lentelių randame, kad $P(Z < -1,64) = 0,05$. Tai reiškia, kad

$$c = 270 - 1,64 \left(\frac{24}{\sqrt{38}} \right) = 270 - 6,39 = 263,61.$$

6. Mūsų sprendimo priėmimo taisyklė šiuo atveju pilnai nusakoma empirinio vidurkio $\bar{X}$ atmetimo sritimi $\mathcal{R} : \bar{X} < 263,61$ (žr. 34 pav.).



34 pav. Nulinės hipotezės atmetimo sritis $\mathcal{R}$

Priimta sakyti, kad nulinės hipotezės patikrinimas atliktas *reikšmingumo lygmenyje*¹ $\alpha = 0,05$, o statistika, kurios pagrindu formuluojama sprendimo priėmimo taisyklė, vadinama *kriterijaus statistika*².

¹ Angl. *significance probability*; rus. *уровень значимости*.

² Angl. *test statistic*. Pavyzdyje tai buvo statistika $\bar{X}$, t.y. $\bar{X}$ yra kriterijaus statistika.

21. DIDELIŲ IMČIŲ NORMALUMO KRITERIJUS. DU KLAIDŲ TIPAI

1. Apibendrinsime pavyzdį, išnagrinėtą praeitame skyrelyje. Tarkime, kad reikia patikrinti nulinę hipotezę

$$H_0 : \mu = \mu_0 \quad \text{prieš alternatyvą} \quad H_1 : \mu < \mu_0.$$

Jei $\bar{X}$ yra normaliųjų atsitiktinių dydžių (su žinoma dispersija σ^2) imties, kurios dydis yra n , empirinis vidurkis, tai *sprendimo priėmimo taisyklė reikšmingumo lygmenyje α yra tokia:*

Atmesti H_0 , jei $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$.

Čia z_α tenkina lygtį $P(Z < -z_\alpha) = P(Z > z_\alpha) = \alpha$ ir yra randamas iš standartinio normaliojo skirstinio uodegų reikšmių 2 lentelės, o μ reikšmės μ_0 parinkimą sąlygoja nulinė hipotezė H_0 .

2. Praeitame skyrelyje išnagrinėtame pavyzdyje su bankomatais tarėme, kad $\sigma = 24$. Atveju $\mu = 270$ matėme, kad $\bar{X}$ skirstinys yra normalusis su parametrais 270 ir $24/\sqrt{38}$, o atmetimo sritis $\mathcal{R} : \bar{X} \leq 263,61$ buvo nustatyta esant reikšmingumo lygmeniui $\alpha = 0,05$ ir turint omeny, kad statistika

$$Z = \frac{\bar{X} - 270}{24/\sqrt{38}}$$

turi standartinį normalųjį skirstinį.

O kaip elgtis tuo atveju, jei dispersija σ yra nežinoma?

Tuo atveju, kai σ nėra žinomas, tačiau n yra pakankamai didelis ($n > 30$), statistikos $\bar{X}$ normalioji aproksimacija išsaugo pakankamai aukštą tikslumo laipsnį net tada, kai σ^2 pakeičiamas empirinės dispersijos reikšme s^2 . Todėl tikrinant $H_0 : \mu = \mu_0$ prieš alternatyvą $H_1 : \mu < \mu_0$ reikšmingumo lygmenyje α kriterijaus statistika pasirenkama statistika

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

o atmetimo sritimi – sritis $\mathcal{R} : Z < -z_\alpha$. Toks pasirinkimas paprastai vadinamas *didelių imčių normalumo kriterijumi*¹.

¹ Angl. *large-sample normal test or a Z-test*.

3. Pavyzdį su banknotais išnagrinėsime tam atvejui, kai σ nėra žinomas, ir parinksime kitą reikšmingumo lygmenį α , $\alpha = 0,025$. Atlikus konkretų eksperimentą nustatyta, kad empirinis vidurkis yra 261 sekundė, o empirinis standartinis nuokrypis yra 22 sekundės. Taigi, reikia patikrinti nulinę hipotezę

$$H_0 : \mu = 270 \text{ prieš alternatyvą } H_1 : \mu < 270.$$

Sprendimas. Kadangi $n = 38$ ir nulinė hipotezė sąlygoja μ reikšmės $\mu_0 = 270$ parinkimą, tai kriterijaus statistika yra

$$Z = \frac{\bar{X} - 270}{S/\sqrt{38}}.$$

Kadangi $\alpha = 0,025$, tai $z_{0,025} = 1,96$ ir atmetimo sritis yra $\mathcal{R} : z \leq -1,96$.

Konkrečios $\bar{X}$ ir S reikšmės yra duotos:

$$\bar{x} = 261, \quad s = 22.$$

Iš čia randame, kad

$$z = \frac{261 - 270}{22/\sqrt{38}} = -2,52,$$

t.y. z atmetimo srityje $\mathcal{R}$, todėl nulinė hipotezė yra atmetama reikšmingumo lygmenyje $\alpha = 0,025$.

4. Suformuluosime hipotezių tikrinimo taisyklę:

HIPOTEZĖS TIKRINIMO ŽINGSNIAI

- 1⁰ Suformuluojama nulinė hipotezė H_0 ir alternatyvioji hipotezė H_1 .
- 2⁰ Norint patikrinti nulinę hipotezę, vartojama specialiai parinkta atsitiktinės imties $X_1, \dots, X_n$ funkcija – atsitiktinis dydis $\mathcal{K} = \mathcal{K}(X_1, \dots, X_n)$, vadinamas *kriterijaus statistika* (mūsų pavyzdžiuose tai buvo $\bar{X}$ arba Z).
- 3⁰ Pasirinkus reikšmingumo lygmenį α , apskaičiuojama kriterijaus statistikos $\mathcal{K}$ atmetimo (arba kritinė) sritis.
- 4⁰ Konkrečiai imčiai $x_1, \dots, x_n$ apskaičiuojama konkreti kriterijaus statistikos $\mathcal{K}$ reikšmė $k = \mathcal{K}(x_1, \dots, x_n)$.
- 5⁰ Pagrindinis statistinės hipotezės tikrinimo principas yra toks: konkrečiai kriterijaus $\mathcal{K}$ reikšmei k patekus į atmetimo (kritinę) sritį, nulinė hipotezė H_0 yra atmetama ir priimama alternatyvioji hipotezė H_1 , o kitu atveju nulinė hipotezė H_0 yra paliekama.

5. Skyrelio pabaigoje pastebėsime, kad tikrindami nulinę hipotezę H_0 galime padaryti šias klaidas:

pirmosios rūšies – kai atmetama teisinga nulinė hipotezė;

antrosios rūšies – kai priimama neteisinga nulinė hipotezė.

Praktiniuose uždaviniuose išskylančios situacijos, taikant sprendimo priėmimo taisyklę ir pagrindinį statistinės hipotezės tikrinimo principą, pavaizduotos šioje lentelėje:

14 lentelė

Sprendimas, priimtas imties pagrindu	Nežinoma reali situacija	
	H_0 yra teisinga	H_1 yra teisinga
H_0 atmetama	Klaidingas H_0 atmetimas (I rūšies klaida)	Teisingas sprendimas
H_0 priimama	Teisingas sprendimas	Klaidingas H_0 priėmimas (II rūšies klaida)

6. Kiti hipotezių tikrinimo klausimai, o būtent:

- paprastoji ir sudėtinė hipotezės, hipotezių tipai,
- statistinis testas, jo kritinė sritis,
- K. Pirsono suderinamumo testas,
- chi kvadrato testo praktinis taikymas,
- Kolmogorovo ir Smirnov testai,

yra gana elementariai nagrinėjami antrojoje vadovėlio dalyje 41-45 skyreliuose.

22. TIESINIS SĄRYŠIS. POSTŪMIO IR KRYPTIES KOEFICIENTAI. TIESINĖS KORELIACIJOS KOEFICIENTAS

1. Realiame gyvenime yra labai daug dydžių, kurie tarpusavyje yra vienaip ar kitaip susiję, t.y. priklausomi. Paprasčiausia iš šių priklausomybių yra tiesinis sąryšis. Šio sąryšio nagrinėjimą pradėsime tokiu pavyzdžiu.

Tarkime, kad kelionė taksi kainuoja 0,9 Lt už vieną pravažiuotą su keleiviu kilometrą. Jei keleivis pravažiavo taksi x km, tai ši jo kelionė kainuos y Lt,

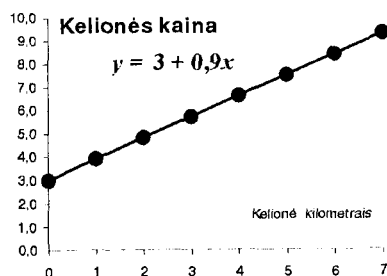
$$y = 0,9x.$$

Tarkime dabar, kad už įlaipinimą imamas 3 Lt papildomas mokestis. Tada, akivaizdu,

$$y = 3 + 0,9x.$$

Konkrečius duomenis nesunku pavaizduoti lentele ir grafiškai (tiesė kyla iš kairės į dešinę – žr. 35 pav.).

Kelionė kilo- metrais x	Kelionės kaina litais y
0	3,0
1	3,9
2	4,8
3	5,7
4	6,6
5	7,5
6	8,4
7	9,3



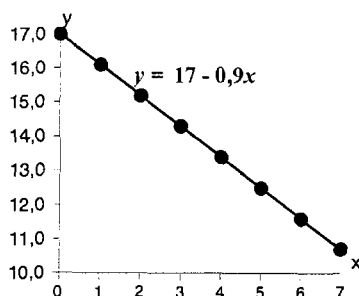
35 pav. Paprasčiausias dydžių tarpusavio priklausomybės atvejis – tiesinis sąryšis

2. Tarkime toliau, kad keleivis, įsėdamas į taksi, turėjo 20 Lt. Tada po x km kelionės jo pinigineje liks

$$y = 20 - (3 + 0,9x) = 17 - 0,9x$$

litų. Konkrečius y priklausomybės nuo x duomenis pateiksime lentele ir grafiškai (tiesė leidžiasi iš kairės į dešinę):

Kelionė kilo- metrais x	Kelionės kaina litais y
0	17,0
1	16,1
2	15,2
3	14,3
4	13,4
5	12,5
6	11,6
7	10,7



36 pav. Tiesinis sąryšis, kuriame krypties koeficientas yra neigiamas

3. Bendru atveju tiesinis sąryšis tarp kintamųjų x ir y matematiškai išreiškiamas tiesine lygtimi

$$y = a + bx,$$

kur koeficientai a ir b yra parametrai, nusakantys tiesės vietą ir kryptį.

Jeigu b , koeficientas prie x , yra teigiamas, tai tiesė kyla iš kairės į dešinę, o jei b yra neigiamas, tai tiesė leidžiasi iš kairės į dešinę. Abi situacijas buvome pavaizdavę grafiškai atvejams $b = 0,9$ ir $b = -0,9$.

Norėdami apskaičiuoti parametą b , suteikime kintamajam x vienetinį prieauglį, t.y. pakeiskime x į $x + 1$. Atitinkamai y pasikeis į y_1 . O kadangi

$$y = a + bx, \quad y_1 = a + b(x + 1),$$

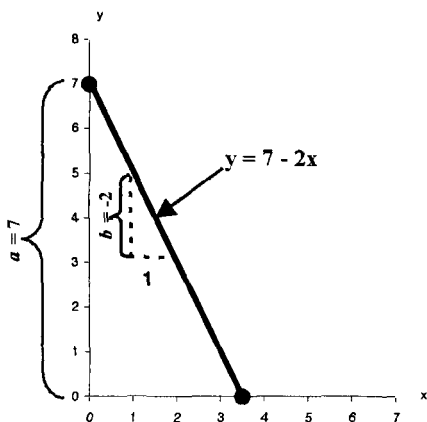
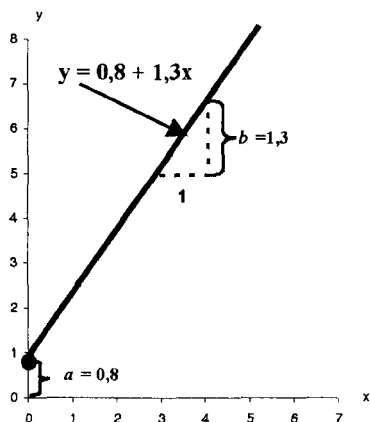
tai atimdami gauname, kad

$$y_1 - y = b.$$

Taigi b reprezentuoja y pokytį, x įgijus vienetinį prieauglį, ir yra vadinamas tiesės *krypties koeficientu*.

Kai $x = 0$, tai $y = a$. Parametras a yra vadinamas *y postūmio koeficientu*.

Postūmio ir krypties koeficientai pavaizduoti šiuose dviejuose grafikuose:



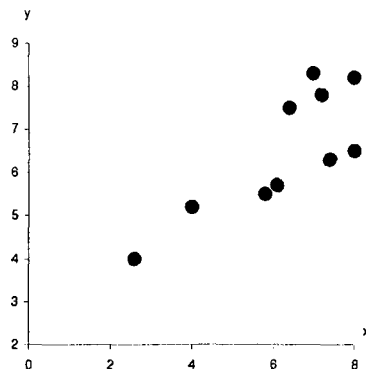
37 pav. Krypties koeficientas reprezentuoja y pokytį, x įgyjus vienetinį prieauglį

4. Reziumuosime tai, kas buvo išdėstyta šiame skyrelyje.

Tiesinio sąryšio lygtis užrašoma forma $y = a + bx$, kurioje postūmio koeficientas a parodo y reikšmę taške $x = 0$, o krypties koeficientas b išreiškia tiesės nuolydį, kuris prilyginamas y pokyčiui, kai x įgyja vienetinį prieauglį. Jei b yra teigiamas, tai y auga, kai auga x ; jei b yra neigiamas, tai y mažėja, kai auga x ; jei, pagaliau, $b = 0$, tai y yra konstanta ir $y = a$.

5. Universitete buvo atliktas tyrimas, kurio tikslas buvo nustatyti, ar egzistuoja priklausomybė (sąryšis) tarp studentų darbo rezultatų viso semestro metu ir egzamino metu. Tuo tikslu buvo atitinkamai pasirinkti 10 studentų. Jų namų darbų rezultatai bei apklausa pratybų metu fiksuojami viso semestro metu, o po to išvestas vidurkis, kuris buvo lyginamas su egzamino rezultatais. Egzamino įvertinimas – tai pažymio už atsakymą žodžiu ir pažymio už uždavinio sprendimą raštu vidurkis. Buvo gauta tokia lentelė ir tokia diagrama:

Pažymiai semestro metu	Pažymiai sesijos metu
5,8	5,5
6,4	7,5
8,0	8,2
7,4	6,3
4,0	5,2
7,0	8,3
7,2	7,8
2,6	4,0
8,0	6,5
6,1	5,7



38 pav. Sąryšio tarp studentų darbo rezultatų viso semestro metu ir egzamino metu tyrimo rezultatai

6. Tokio tipo diagramos dar yra vadinamos *sklaidos diagramomis*¹. Išanalizavę šią diagramą ir duomenis lentelėje darome išvadą, kad tie studentai, kurie gavo aukštesnius pažymius semestro metu, gavo aukštus pažymius ir egzamino metu. Atvirkščiai, tie studentai, kurie gavo mažus pažymius semestro metu, gavo mažus pažymius ir egzamino metu. Tokio tipo sąryšis tarp x ir y vadinamas *teigiamu sąryšiu*².

¹ Angl. *scatter diagram*, rus. *векторная диаграмма*.

² Angl. *positive relationship*.

Iš kitos pusės, jei didelėms x reikšmėms atitinka mažos y reikšmės, o mažoms x reikšmėms atitinka didelės y reikšmės, tai toks sąryšis tarp x ir y vadinamas *neigiamu* (arba *atvirkštinu*) *sąryšiu*¹. Neigiamo sąryšio pavyzdžiu galėtų būti sąryšis tarp automobilio kainos ir jo eksploatacijos metų skaičiaus.

7. Iš diagramos matome, kad taškai (x_i, y_i) išdėstyti ne atsitiktinai, ne bet kaip, o išsisklaidę tam tikru būdu aplink tiesę, daugiau ar mažiau su ja susieti nagrinėjamo sąryšiu.

Matas, charakterizuojantis sąryšio artumą šiai tiesei, yra vadinamas tiesinės koreliacijos koeficientu arba tiesiog koreliacijos koeficientu:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

23. KORELIACIJOS KOEFICIENTO APSKAIČIAVIMAS IR JO SAVYBĖS. REGRESIJOS TIESĖS PARINKIMAS

1. Koreliacijos koeficiento praktiniam skaičiavimui yra patogi tokia taisyklė.

Taškų $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ koreliacijos koeficientu yra vadinamas skaičius r ,

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\{n \sum x_i^2 - (\sum x_i)^2\} \{n \sum y_i^2 - (\sum y_i)^2\}}}.$$

Norint rasti r skaitinę išraišką reikia, tokiu būdu, rasti skaitines išraiškas $\sum x_i \sum y_i$, $\sum x_i^2$, $\sum y_i^2$ ir $\sum x_i y_i$. Daugelis kalkuliatorių turi šias funkcijas, o kai kuriais iš jų koreliacijos koeficientą r galima apskaičiuoti ir tiesiogiai.

2. Apskaičiuosime koreliacijos koeficientą praeitame skyrelyje išnagrinėtame pavyzdyje apie sąryšį tarp studentų darbo rezultatų viso semestro metu ir egzamino metu. Šiame pavyzdyje $n = 10$ ir buvo gauta 10 taškų porų: (5,8; 5,5); (6,4; 7,5); (8,0; 8,2); (7,4; 6,3); (4,0; 5,2); (7,0; 8,3); (7,2; 7,8); (2,6; 4,0); (8,0; 6,5); (6,1; 5,7). Sudarome 15 lentelę.

Iš koreliacijos koeficiento r formulės gauname, kad

$$r = \frac{10 \cdot 424,35 - 62 \cdot 5 \cdot 65}{\sqrt{(10 \cdot 418,17 - (62,5)^2)(10 \cdot 440,94 - (65)^2)}} = 0,803.$$

¹ Angl. *negative (inverse) relationship*.

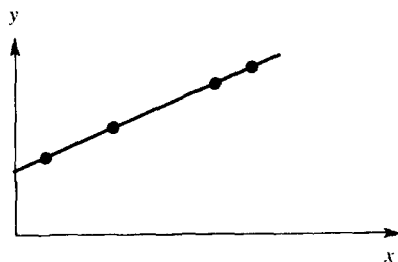
15 lentelė

X_i	y_i	x_i^2	y_i^2	$x_i y_i$
5,8	5,5	33,64	30,25	31,90
6,4	7,5	40,96	56,25	48,00
8,0	8,2	64,00	67,24	65,60
7,4	6,3	54,76	39,69	46,62
4,0	5,2	16,00	27,04	20,80
7,0	8,3	49,00	68,89	58,10
7,2	7,8	51,84	60,84	56,16
2,6	4,0	6,76	16,00	10,40
8,0	6,5	64,00	42,25	52,00
6,1	5,7	37,21	32,49	34,77
Suma: 62,5	65,0	418,17	440,94	424,35

3. Korelācijas koeficienta savybės:

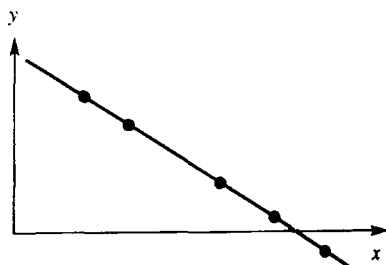
1⁰. Korelācijas koeficienta reikšmės yra tarp -1 ir 1, t.y. $-1 \leq r \leq 1$.

2⁰. Atvejis $r = 1$ reiškia, kad tarp y ir x yra tikslus teigiamas tiesinis sąryšis, o jos sklaidos diagramoje (39 pav. (a)) visi taškai išdėstyti tiesėje, kuri kyla iš kairės į dešinę.



39 pav. (a) $r = 1$, teigiamas tiesinis sąryšis

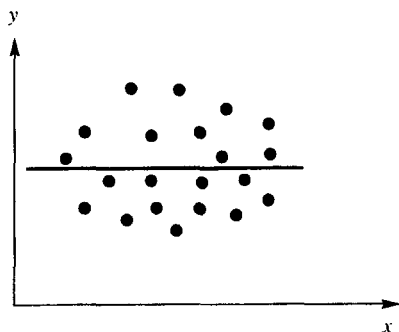
3⁰. Atvejis $r = -1$ reiškia, kad tarp y ir x yra tikslus neigiamas tiesinis sąryšis, o jos sklaidos diagramoje (39 pav. (b)) visi taškai išdėstyti tiesėje, kuri leidžiasi iš kairės į dešinę.



39 pav. (b) $r = -1$, neigiamas tiesinis sąryšis

4⁰. Atvejis, kai r yra arti nulio, reiškia, kad yra galimas vienas iš dviejų variantų:

4.1⁰. Nagrinėjami dydžiai nėra arba beveik nėra susiję. Paveiksliuke 39 (c) pavaizduota nepriklausomų dydžių sklaidos diagrama. Tokiu atveju sąryšis tarp x ir y gali būti pavaizduotas horizontalia tiese. Dydžiai beveik nėra susiję.

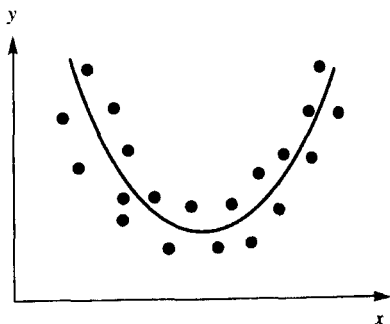
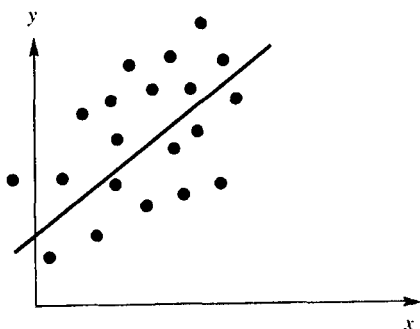
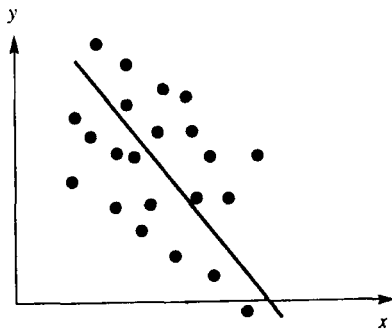


39 pav. (c) r arti nulio. Dydžiai beveik nėra susiję

4.2⁰. Nagrinėjami dydžiai yra susiję, bet ne tiesiniu sąryšiu, o, pavyzdžiui, taip, kaip yra pavaizduota 39 pav. (d). Šiame piešinyje r yra arti nulio.

5⁰. Jei r yra teigiamas, tai *tiesinis trendas*¹ kyla iš kairės į dešinę, o jei neigiamas, tai leidžiasi (žr. 39 pav. (e) ir (f)).

¹ Angl. *trend* – tendencija, kryptis, linkmė.

39 pav. (d) r arti nulio. Dydžiai nėra susiję tiesiniu sąryšiu39 pav. (e) $0 < r < 1$, sąryšis teigiamas39 pav. (f) $-1 < r < 0$, sąryšis neigiamas

4. Regresijos tiesės parinkimas

Žinodami, kad $|r| = 1$, mes galime tvirtinti, kad abu nagrinėjami dydžiai yra susiję tiesiniu sąryšiu. Dar daugiau, jei $|r|$ yra artimas vienetui, tai nagrinėjamas sąryšis yra artimas tiesiniam. Tačiau koreliacijos koeficientas r neteikia informacijos, koks yra tas tiesinis sąryšis, kokie yra postūmio ir krypties koeficientai. Šie klausimai yra nagrinėjami regresinėje analizėje.

Joje yra priimta nepriklausomą kintamąjį žymėti raide x , o jo reikšmės grafike atidėti horizontaliojoje x ašyje. Priklausomas kintamasis žymimas raide y , o jo reikšmės, kaip įprasta, fiksuojamos vertikalioje ašyje.

Išnagrinėjome tyrimo rezultatus apie priklausomybę tarp studentų darbo rezultatų viso semestro metu ir egzamino metu. Šio tyrimo sklaidos diagrama

pateikta 38 pav. Ši diagrama leidžia daryti išvadą, kad nagrinėjamas sąryšis yra artimas tiesiniam. Jei visi taškai būtų išdėstyti vienoje tiesėje, rasti šios tiesės pastūmio ir krypties koeficientus nebūtų sudėtinga, nes jie nusakomi vienareikšmiškai šios tiesės vieta plokštumoje. Tačiau nagrinėjamu atveju taškai nėra sukoncentruoti vienoje tiesėje, o per juos pačius teoriškai galima nubrėžti be galo daug tiesių.

Taigi, iškyla toks uždavinys. Nubrėžti tiesę, kuri sklaidos diagramoje *geriausiai* atspindėtų taškų išdėstymo tiesiškumo tendenciją, t.y. *tiesinį trendą*.

Tačiau ką reiškia geriausiai? Detalizuodami šią sąvoką ir patį principą, leidžiantį nubrėžti tiesę, kuri geriausiai atspindėtų taškų išdėstymą, mes vadosimės Ležandro¹ *mažiausių kvadratų metodu*.

24. PROGNOZĖS LYGTIS. MAŽIAUSIŲ KVADRATŲ METODAS. REGRESIJOS KOEFICIENTAS

1. Pakartosime uždavinio, kurį spęsimė šiame skyrelyje, formulavimą. Nubrėžti tiesę, kuri sklaidos diagramoje geriausiai atspindėtų taškų išdėstymo tiesiškumo tendenciją, t.y. *tiesinį trendą*.

Tarkime, kad $\hat{y} = a + bx$ ir yra ši tiesė, t.y. tiesė $\hat{y} = a + bx$ geriausiai „išlygina“ n porų reikšmių $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Ši tiesinė lygtis ir yra vadinama *prognozės lygtimi*. Sakoma, kad taškas $\hat{y}_i = a + bx_i$ reprezentuoja kintamojo y prognozuojamą reikšmę taške x_i . Suprantama, kad y_i yra atitinkama reali (o ne prognozuojama) kintamojo y reikšmė taške x_i .

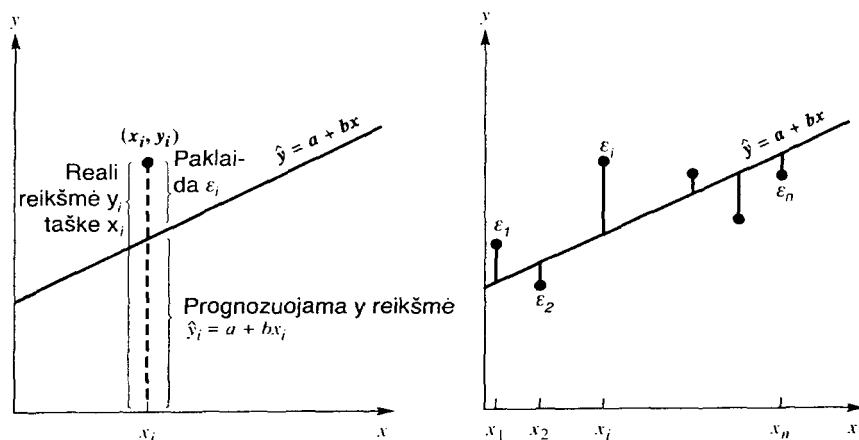
2. Dydžiai $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n$, t.y. $y_1 - (a + bx_1), y_2 - (a + bx_2), \dots, y_n - (a + bx_n)$ – tai stebimų (realių) kintamojo y reikšmių nuokrypiai nuo prognozuojamų reikšmių ir kiekvienas iš jų vadinamas *netiktimi* (liekana)² arba *paklaida*³ ir žymima $\varepsilon_i, \varepsilon_i = y_i - \hat{y}_i$. Grafiškai tai pavaizduota 40 pav.

3. Iš grafiko matome, kad taškai $\varepsilon_i = y_i - (a + bx_i)$ – tai vertikalūs atstumas nuo stebimo (realaus) y_i iki prognozės tiesės, kurios parametrus a ir b mums dar reikia apibrėžti. Šie koeficientai randami mažiausių kvadratų metodo, kurį tuoj suformuluosime, pagrindu.

¹ Andrien-Marie Legendre (1752-1833).

² Angl. *residual*; rus. *невязка*.

³ Angl. *error*, rus. *погрешность*.



40 pav. Tiesės, kuri sklaidos diagramoje geriausiai atspindi tiesinį trendą, radimas

Mažiausių kvadratų metodo tikslas:

Iš aibės galimų tiesių sklaidos diagramoje išrinkti geriausių tiesę ta prasme, kad paklaidų kvadratų suma būtų minimali. Tai yra, apibrėžti koeficientus a ir b taip, kad suma

$$\sum \epsilon_i^2 = \sum (y_i - (a + bx_i))^2$$

būtų minimali.

4. Tiesė, gauta naudojant mažiausių kvadratų metodą, dar yra vadinama mažiausių kvadratų regresijos tiese. Naudojant skaičiavimo metodus apskaičiuojami patys koeficientai a ir b , minimizuojantys paklaidų kvadratų sumą.

Mažiausių kvadratų formulės:

Mažiausių kvadratų regresijos tiesės krypties koeficientas b apibrėžiamas formule

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2},$$

o pastūmio koeficientas a – formule

$$a = \bar{y} - b\bar{x},$$

kur n yra stebėjimų $(x_1, y_1), \dots, (x_n, y_n)$ skaičius, o $\bar{y}$ yra reikšmių $y_1, \dots, y_n$ empirinis vidurkis, $\bar{x}$ – reikšmių $x_1, \dots, x_n$ empirinis vidurkis.

Krypties koeficientas b , gautas mažiausių kvadratų metodu, dar yra vadinamas *regresijos koeficientu*, o pati tiesė $\hat{y} = a + bx$, gauta šiuo metodu, sutrumpintai vadinama *regresijos tiese*.

5. Pavyzdys. Rasime regresijos koeficientą ir nubrėšime regresijos tiesę uždavinyje apie sąryšį tarp studentų darbo rezultatų viso semestro metu ir egzamino metu. Ši problema detalai suformuluota ir pradėta nagrinėti 22-ame ir 23-ame skyreliuose.

Sprendimas. Pirmiausia apskaičiuosime krypties koeficientą b , o po to ir postūmio koeficientą a . Dydžiai, kurie šiam tikslui mums reikalingi, yra šie: $\sum x_i y_i$, $\sum x_i^2$, $\sum x_i$, $\sum y_i$. Visi jie yra 15 lentelėje 23-ame skyrelyje, iš kurios randame, kad $n = 10$,

$$\sum_{i=1}^{10} x_i y_i = 424,35; \quad \sum_{i=1}^{10} x_i^2 = 418,17; \quad \sum_{i=1}^{10} x_i = 62,5; \quad \sum_{i=1}^{10} y_i = 65,0.$$

Iš mažiausių kvadratų formulės gauname, kad

$$b = \frac{10 \cdot 424,35 - 62,5 \cdot 65}{10 \cdot 418,17 - 62,5^2} = \frac{181}{275,45} = 0,657.$$

Kadangi $\bar{x} = 62,5/10$, o $\bar{y} = 65/10$, tai

$$a = \bar{y} - b\bar{x} = 6,5 - 0,657 \cdot 6,25 = 2,394.$$

Taigi, regresijos tiesė, o tuo pačiu ir prognozės lygtis yra ši:

$$\hat{y} = a + bx = 2,394 + 0,657x.$$

Norint nubrėžti regresijos tiesę sklaidos diagramoje (žr. 41 pav.), pakanka žinoti bent du jos taškus:

(a) Jei $x = 5,8$, tai $\hat{y} = 2,394 + 0,657 \cdot 5,8 = 6,205$,

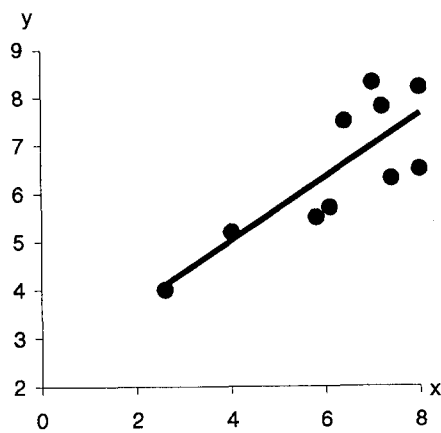
(b) Jei $x = 8,0$, tai $\hat{y} = 2,394 + 0,658 \cdot 8,0 = 7,658$.

6. Pastaba. Iš mažiausių kvadratų formulės postūmio koeficientui a turime, kad

$$a = \bar{y} - b\bar{x}, \quad \text{t.y.} \quad \bar{y} = a + b\bar{x}.$$

Tai, savo ruožtu, reiškia, kad *regresijos tiesei visada priklauso taškas* $(\bar{x}, \bar{y})$.

Mūsų pavyzdyje matome, kad regresijos tiesė kerta tašką $(\bar{x}, \bar{y}) = (6,25; 6,5)$.



41 pav. Regresijos tiesė, gauta mažiausių kvadratų metodu

II. MATEMATINIAI STATISTIKOS METODAI

„Per pastaruosius dvidešimt penkerius metus statistika kaip mokslas stipriai progresavo. Šis progresas pasiektas puikiais anglų ir amerikiečių statistikos mokyklų atstovų darbais, kurių tarpe pirmiausiai turi būti paminėti profesoriaus R.A. Fišerio (Fisher) darbai. Šiuo laikotarpiu daugiausia prancūzų ir rusų matematikų dėka klasikinis tikimybių skaičiavimas virto grynai matematine teorija, tenkinančia šiuolaikinius griežtus reikalavimus“ – taip dar 1945 metais rašė žymus Stokholmo universiteto profesorius H. Krameras savo knygoje [7].

Būtent įvadui į šią matematinę teoriją, jos šiuolaikinius metodus ir skirta antroji mūsų vadovėlio dalis.

Kai kurie terminai apibrėžiami du kartus: 1-oje ir šioje dalyje. Taip daroma Lietuvos standarto [8] pavyzdžiu, norint atkreipti skaitytojo dėmesį į tai, kad

a) šių terminų **tikimybinių interpretacija** remiasi teorinėmis konstrukcijomis, kurios nesusijusios su praktiniais taikymais;

b) šių terminų **statistinė interpretacija** remiasi *stebinio* sąvoka, o terminų apibrėžimai yra taikomojo pobūdžio.

25. DISKREČIOJI ELEMENTARIŲJŲ ĮVYKIŲ ERDVĖ IR JOS MATO SAVYBĖS

1. Antrajame skyrelyje stochastinio eksperimento visų galimų rezultatų aibę Ω pavadiname *elementariųjų įvykių erdve*, o jos elementus ω_j – *elementariaisiais įvykiais* arba *elementariosiomis baigtimis*.

Dabar detaliau išnagrinėsime stochastinį eksperimentą su baigtine arba skaičiąja visų baigčių $\omega_1, \dots, \omega_i, \dots$ aibe Ω , t.y. $\Omega = \{\omega_1, \dots, \omega_i, \dots\}$. Tokia aibė Ω vadinama *diskrečiąja elementariųjų įvykių erdve*.

Tarkime, kad kiekvienam elementariajam įvykiui ω_i priskirtas „svoris“ p_i (kurį pavadinsime elementariojo įvykio ω_i *tikimybe*) ir kad šie svoriai turi tokias savybes:

$$1) p_i \geq 0,$$

$$2) \sum_{i=1}^{\infty} p_i = 1.$$

2. Tegu A – bet koks atsitiktinis įvykis, stebimas nagrinėjamame stochastiniame eksperimente, $A \subseteq \Omega$.

Atsitiktinio įvykio A *tikimybe* $P(A)$ vadiname *elementariųjų įvykių, su-*

darančių įvykį A , tikimybių suma:

$$P(A) = \sum_{i: \omega_i \in A} p_i. \quad (1)$$

Tokiu būdu įvesta tikimybė turi savybes, kurias suformuluosime kaip atskirą teoremą.

Teorema. Tegu Ω yra diskrečioji elementariųjų įvykių erdvė, kurioje tikimybė $P(A)$ yra apibrėžiama formule (1). Tada

$$(a) 0 \leq P(A) \leq 1,$$

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

$$(c) \text{ jei } A \cap B = \emptyset, A \subseteq \Omega, B \subseteq \Omega, \text{ tai } P(A \cup B) = P(A) + P(B).$$

Irodymas. Teoremos išvados (a) ir (b) yra elementarios, nes iš $\{p_i\}$ savybių 1) ir 2) gauname, kad

$$0 \leq P(A) = \sum_{i: \omega_i \in A} p_i \leq \sum_{i=1}^{\infty} p_i = 1,$$

$$P(\Omega) = \sum_{i: \omega_i \in \Omega} p_i = \sum_{i=1}^{\infty} p_i = 1.$$

Kadangi $A \cap B = \emptyset$, tai

$$P(A \cup B) = \sum_{i: \omega_i \in A \cup B} p_i = \sum_{i: \omega_i \in A} p_i + \sum_{i: \omega_i \in B} p_i = P(A) + P(B).$$

Teorema įrodyta.

3. Diskrečiosios elementariųjų įvykių erdvės Ω poaibių $\{A\}$ funkcija $P(\cdot)$, turinti savybes (a), (b) ir (c), dar yra vadinama *tikimybiniu matu*, o pats dvejetas (Ω, P) – *diskrečiuoju tikimybiniu modeliu*.

Kaip „teisingai“ apibrėžti elementariųjų įvykių ω_i tikimybes p_i ? Šis uždavinys nepriklauso tikimybių teorijos sričiai. Nežinomų tikimybių įvertinimo metodai pagal stebėjimų rezultatus nagrinėjami matematinėje statistikoje.

Šiame skyrelyje tikimybiniis matas apibrėžtas teorinės konstrukcijos pagrindu ir matematiškai griežtai – kaip aibių funkcija. Koks gi šio apibrėžimo intuityvus pagrindimas? Ši tikimybės apibrėžimą galima sieti su įsitikinimo, kad įvykis įvyks, laipsniu. Kai įsitikinimo laipsnis didelis, pastebima standarte [8], tai tikimybė artima 1.

Kitas tikimybinio mato įvedimas išnagrinėtas 4-ajame skyrelyje. Jis buvo grindžiamas statistinio stabilumo fenomenu, o įvykio tikimybė apibrėžiama kaip santykinio dažnio riba.

Taigi *tikimybę* apibrėžti galima dviem būdais: kaip santykinio dažnio ribą ir kaip tikėtumo laipsnį. Beje, pastarajam būdai priskirtinas ir taip vadinamas **klasikinis tikimybės apibrėžimas**. Jei *stochastiniame eksperimente yra tik n vienodai galimų baigčių, t.y. jei diskrečioji elementariųjų įvykių erdvė sudaryta iš n vienodai galimų elementariųjų įvykių, tai įvykio A tikimybė vadinamas santykiu*

$$P(A) = m/n,$$

kur m – elementariųjų įvykių, palankių įvykiui A , skaičius.

4. Pavyzdžiai. Stochastinis eksperimentas yra toks – metamas simetrinis kubiukas. Tada diskrečiąją elementariąją įvykių erdvę Ω natūralu apibrėžti taip:

$$\Omega = \{1, 2, 3, 4, 5, 6\} = \{\omega_i \mid \omega_i = i, i = 1, 2, 3, 4, 5, 6\}.$$

Kadangi kubiukas yra simetrinis, tai kiekvienam elementariajam įvykiui $\omega_i = i$ priskirsime tikimybę $p_i = 1/6$ ($i = 1, 2, \dots, 6$). Tai ir yra diskretusis tikimybės modelis nagrinėjamam stochastiniam eksperimentui.

Tegu A yra įvykis, kuris reiškia, kad iškritęs skaičius dalijasi iš 3. Tada $A = \{\omega_3, \omega_6\}$,

$$P(A) = \frac{m}{n} = \frac{2}{6} = \frac{1}{3}.$$

Dabar tarkime, kad atliekamas toks stochastinis eksperimentas – metamas nesimetrinis kubiukas. Yra žinoma, kad jo sienelių masė proporcinga sienelės numeriui.

Diskretųjį tikimybinį modelį konstruosime taip. Diskrečioji elementariųjų įvykių erdvė Ω apibrėžiama taip pat, kaip ir pirmajame pavyzdyje. Kiekvienam elementariajam įvykiui $\omega_i = i$ priskirsime tikimybę p_i ,

$$p_i = i / \sum_{k=1}^6 k = i/21.$$

Lengva įsitikinti, kad taip apibrėžtas skaičių rinkinys $\{p_i\}$ tenkina savybes 1) ir 2). Įvykio $A = \{\omega_3, \omega_6\}$ tikimybė, aišku, skirsis nuo šio įvykio tikimybės, apskaičiuotos pirmajame pavyzdyje, nes kubiukas nėra simetrinis. Taigi,

$$P(A) = P(\{\omega_3\}) + P(\{\omega_6\}) = p_3 + p_6 = \frac{3}{21} + \frac{6}{21} = \frac{3}{7}.$$

26. AIBIŲ ALGEBRA IR σ -ALGEBRA. TEOREMA APIE MINIMALIOS σ -ALGEBROS EGZISTAVIMĄ

1. Aibių algebros ir σ -algebros sąvokų įvedimas ypač aktualus tada, kai stochastinio eksperimento visų galimų rezultatų aibė Ω nėra diskrečioji elementariųjų įvykių erdvė, t.y. Ω yra neskaičioji aibė.

Priminsime, kad aibės, ekvivalenčios natūrinių skaičių aibei, pasižymi tuo, kad jų elementus galima tam tikru būdu numeruoti, skaičiuoti. Tokios aibės vadinamos *skaičiosiomis*¹. Begalinės aibės, neekvivalenčios natūrinių skaičių aibei, vadinamos *neskaičiosiomis*.

2. Priminsime kitą apibrėžimą.

Aibės Ω poaibių klasė $\mathcal{A}$ vadinama algebra, jei ji turi tokias savybes:

1⁰. $\Omega \in \mathcal{A}$;

2⁰. Jei $A \in \mathcal{A}$, tai $\bar{A} \in \mathcal{A}$;

3⁰. Jei $A \in \mathcal{A}$ ir $B \in \mathcal{A}$, tai $A \cup B \in \mathcal{A}$.

Teiginys 1. *Tegu $\mathcal{A}$ yra algebra. Jei $A \in \mathcal{A}$, $B \in \mathcal{A}$, tai $A \cap B \in \mathcal{A}$.*

Įrodymas. Pirmiausia įrodysime, kad

$$A \cap B = \overline{\bar{A} \cup \bar{B}}. \quad (1)$$

Tai nesunku atlikti, prisiminus de Morgano logikos dėsnį, pagal kurį bet kuriems teiginiams p ir q yra teisingas toks sąryšis:

$$\overline{p \wedge q} \equiv \bar{p} \vee \bar{q}.$$

Taigi,

$$\begin{aligned} \omega \in A \cap B &\iff (\omega \in A) \wedge (\omega \in B) \iff \overline{(\omega \notin A) \vee (\omega \notin B)} \\ &\iff \overline{(\omega \in \bar{A}) \vee (\omega \in \bar{B})} \iff \overline{\omega \in (\bar{A} \cup \bar{B})} \\ &\iff \omega \notin (\bar{A} \cup \bar{B}) \iff \omega \in \overline{(\bar{A} \cup \bar{B})}, \end{aligned}$$

o tuo pačiu įrodyta formulė (1).

Iš 2⁰ savybės gauname, kad $\bar{A} \in \mathcal{A}$, $\bar{B} \in \mathcal{A}$, o iš 3⁰ – kad $\bar{A} \cup \bar{B} \in \mathcal{A}$. Vėl pasinaudoję savybe 2⁰ turime, kad

$$\overline{\bar{A} \cup \bar{B}} \in \mathcal{A}.$$

¹ Angl. *countable*, rus. *сечетное*.

Iš čia ir iš formulės (1) darome išvadą, kad

$$A \cap B \in \mathcal{A}.$$

Teiginys įrodytas.

3. Matyt, paprasčiausias netrivialus algebros pavyzdys yra toks. Tegu $A \subset \Omega$. Tada aibių klasė $\mathcal{A}$

$$\mathcal{A} = \{\Omega, \emptyset, A, \bar{A}\}$$

yra algebra. Siūlome tai skaitytojui įrodyti savarankiškai.

Teiginys 2. Tegu Ω yra intervalas $[0; 1)$. Jei A yra tokių poaibių iš Ω sistema, kurių kiekvienas yra nesikertančių intervalų tipo $[a; b)$ baigtinė suma, t.y.

$$\mathcal{A} = \left\{ \bigcup_{j=1}^n [a_j; b_j) \mid 0 \leq a_1 < b_1 \leq a_2 < b_2 \leq \dots \leq a_n < b_n \leq 1, n < \infty \right\}, \quad (2)$$

tai $\mathcal{A}$ yra algebra.

Įrodymas. Pirmiausia pastebėsime, kad $\Omega \in \mathcal{A}$, o formulę (1) užrašysime bendresniu pavidalu:

$$\overline{\bigcup_{j=1}^n A_j} = \bigcap_{j=1}^n \bar{A}_j. \quad (3)$$

Tegu dabar A – bet koks $\mathcal{A}$ elementas. Ar $\bar{A} \in \mathcal{A}$, t.y. ar nagrinėjamoji sistema $\mathcal{A}$ tenkina savybę 2^0 ?

Jei $A \in \mathcal{A}$, tai

$$A = \bigcup_{j=1}^k [a_j; b_j).$$

Pažymėkime $A_j = [a_j; b_j)$. Tada iš (3) gauname, kad

$$\bar{A} = \overline{\left(\bigcup_{j=1}^k A_j \right)} = \bigcap_{j=1}^k \bar{A}_j = \bigcap_{j=1}^k ([0; a_j) \cup [b_j; 1)).$$

Žymėjimų supaprastinimui toliau tarkime, kad $k = 2$. Tada

$$\bar{A} = ([0; a_1) \cup [b_1; 1)) \cap ([0; a_2) \cup [b_2; 1)).$$

Kadangi $a_1 < b_1 \leq a_2 < b_2$, tai pastebime, kad

$$\bar{A} = [0; a_1) \cup [b_1; a_2) \cup [b_2; 1),$$

t.y. $\bar{A}$ yra baigtinė nesikertančių intervalų tipo $[a; b)$ suma ir todėl $\bar{A} \in \mathcal{A}$.

Dabar belieka įrodyti, kad

$$A \in \mathcal{A}, C \in \mathcal{A} \Rightarrow A \cup C \in \mathcal{A},$$

t.y. nagrinėjamoji sistema $\mathcal{A}$ tenkina ir savybę σ^0 . Taigi, tegu

$$A = \bigcup_{j_1=1}^{k_1} [a_{j_1}; b_{j_1}), \quad C = \bigcup_{j_2=1}^{k_2} [c_{j_2}; d_{j_2}).$$

Ar šių aibių suma

$$A \cup C = \left\{ \bigcup_{j_1=1}^{k_1} [a_{j_1}; b_{j_1}) \right\} \cup \left\{ \bigcup_{j_2=1}^{k_2} [c_{j_2}; d_{j_2}) \right\} \quad (4)$$

turi (2) pavidalą? Čia, pasirinkus bet kuriuos i ir j , galimi tik du variantai:

(I). $[a_i; b_i) \cap [c_j; d_j) = \emptyset$. Tada sumoje (4) paprasčiausiai atliekama nauja pernumeracija.

(II). $[a_i; b_i) \cap [c_j; d_j) \neq \emptyset$, t.y. intervalai $[a_i; b_i)$ ir $[c_j; d_j)$ turi bendrą tašką. Tai reiškia, kad arba $b_i > c_j$, arba $d_j > a_i$. Pirmuoju atveju

$$[a_i; b_i) \cup [c_j; d_j) = [\min(a_i; c_j); \max(b_i; d_j)),$$

o antruoju –

$$[c_j; d_j) \cup [a_i; b_i) = [\min(c_j; a_i); \max(d_j; b_i)),$$

t.y. tas pats. Taigi, $A \cup C$ galima išdėstyti kaip baigtinę nesikertančių intervalų $[\cdot; \cdot)$ sumą, kadangi visus susikertančius intervaliukus apjungėme į to paties tipo intervalą. Tai reiškia, kad $A \cup C \in \mathcal{A}$.

Teiginys įrodytas.

4. Apibrėždami algebrą reikalavome, kad bet kurių jos elementų baigtinė suma jai priklausytų vėl. Tačiau yra daug tikimybių teorijos ir matematinės statistikos uždavinių, kuriuos sprendžiant tenka nagrinėti begalines sumas. Todėl ir tikimybiniis matas įvedamas ne algebroje, o σ -algebroje, kurios apibrėžimą dabar ir pateiksime.

Aibės Ω poaibių klasė $\mathcal{F}$ vadinama σ -algebra, jei ji turi tokias savybes:

1⁰. $\Omega \in \mathcal{F}$;

2⁰. Jei $A \in \mathcal{F}$, tai $\bar{A} \in \mathcal{F}$;

3⁰. Jei $A_j \in \mathcal{F}$, $j = 1, 2, \dots$, tai $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

Skaitytojui siūlome įsitikinti, kad aibė visų Ω poaibių sudaro σ -algebrą.

5. Toliau mums yra reikalingas dar vienas apibrėžimas.

Tegu $\mathcal{K}$ – bet kuri aibės Ω poaibių klasė. σ -algebra $\sigma(\mathcal{K})$ vadinama *mažiausia σ -algebra, talpinančia klasę $\mathcal{K}$* , jei ji tenkina šias savybes:

1⁰. $\mathcal{K} \subset \sigma(\mathcal{K})$.

2⁰. Jei $\mathcal{F}$ – bet kuri σ -algebra tokia, kad $\mathcal{K} \subset \mathcal{F}$, tai $\sigma(\mathcal{K}) \subset \mathcal{F}$.

Teorema. *Bet kuriai klasei $\mathcal{K}$ egzistuoja mažiausia σ -algebra, talpinanti klasę $\mathcal{K}$.*

Įrodymas. Pirmiausia pastebėsime, kad egzistuoja bent viena σ -algebra, talpinanti klasę $\mathcal{K}$. Iš tiesų, kadangi $\mathcal{K}$ yra aibės Ω poaibių klasė, tai tokia σ -algebra yra, pavyzdžiui, visų Ω poaibių σ -algebra.

Visų σ -algebų, talpinančių klasę $\mathcal{K}$, pjūvį pažymėkime $\sigma(\mathcal{K})$. Naudojant vien tik σ -algebros apibrėžimą siūlome skaitytojui savarankiškai patikrinti, kad $\sigma(\mathcal{K})$ yra σ -algebra.

Belieka pastebėti, kad jei $\mathcal{F}$ – bet kuri σ -algebra, talpinanti $\mathcal{K}$, tai pagal mūsų konstrukciją

$$\sigma(\mathcal{K}) \subset \mathcal{F}.$$

O tai reiškia, kad $\sigma(\mathcal{K})$ yra mažiausia σ -algebra, talpinanti klasę $\mathcal{K}$.

Teorema įrodyta.

27. BORELIO σ -ALGEBRA. TIKIMYBINIS MATAS MAČIOJOJE ERDVĖJE

1. Pracitame skyrelyje daug dėmesio skyrėme intervalų $[a; b)$ klasės nagrinėjimui. Šiame skyrelyje tarę, kad $\Omega = \mathbb{R}^1 = (-\infty, \infty)$, šių intervalų $[a; b)$ klasę pažymėkime raide $\mathcal{K}$. *Mažiausia σ -algebra, talpinanti klasę $\mathcal{K}$, vadinama Borelio σ -algebra¹, jos elementai vadinami Borelio aibėmis, o žymima ji raide $\mathcal{B}$.*

¹ Analogiškai galima apibrėžti Borelio σ -algebrą erdvėje $\mathbb{R}^n$.

Teorema. Vienataškė aibė $\{a\}$, atviras intervalas $(a; b)$ ir uždaras intervalas $[a; b]$ yra Borelio aibės.

Įrodymas. Pirmiausia įrodysime, kad jei $\mathcal{A}$ yra bet kokia σ -algebra ir $A_j \in \mathcal{A}$, $j = 1, 2, \dots, n, \dots$, tai

$$\bigcap_{j=1}^{\infty} A_j \in \mathcal{A}. \quad (1)$$

Kadangi

$$\begin{aligned} \omega \in \bigcap_{j=1}^{\infty} A_j &\iff \omega \in A_j, \forall j \iff \omega \notin \bar{A}_j, \forall j \\ &\iff \omega \notin \bigcup_{j=1}^{\infty} \bar{A}_j \iff \omega \in \overline{\bigcup_{j=1}^{\infty} \bar{A}_j}. \end{aligned}$$

Tai reiškia, kad

$$\bigcap_{j=1}^{\infty} A_j = \overline{\bigcup_{j=1}^{\infty} \bar{A}_j}.$$

O kadangi $\mathcal{A}$ yra σ -algebra, tai

$$\begin{aligned} A_j \in \mathcal{A} \Rightarrow \bar{A}_j \in \mathcal{A} \Rightarrow \bigcup_{j=1}^{\infty} \bar{A}_j \in \mathcal{A} \Rightarrow \\ \Rightarrow \overline{\bigcup_{j=1}^{\infty} \bar{A}_j} \in \mathcal{A} \Rightarrow \bigcap_{j=1}^{\infty} A_j \in \mathcal{A}. \end{aligned}$$

Formulė (1) pilnai įrodyta.

Kadangi

$$\{a\} = \bigcap_{n=1}^{\infty} [a; a + 1/n),$$

o aibės $[a; a + 1/n)$ yra Borelio σ -algebros $\mathcal{B}$ elementai, tai iš (1) gauname, kad $\{a\} \in \mathcal{B}$. Pastebėsime, kad jei $B \subset A$, tai $A \setminus B = A \cap \bar{B}$. Iš čia gauname, kad jei $A \in \mathcal{A}$, $B \in \mathcal{A}$, $\mathcal{A}$ – bet kokia σ -algebra, tai $A \setminus B \in \mathcal{A}$. O kadangi

$$(a; b) = [a; b) \setminus \{a\},$$

tai $(a; b) \in \mathcal{B}$. Dabar belieka pastebėti, kad

$$[a; b] = [a; b) \cup \{b\} \in \mathcal{B}.$$

Iš tiesų, $[a; b) \in \mathcal{K}$, taigi $[a; b) \in \mathcal{B}$, nes $\mathcal{K} \subset \mathcal{B}$. Be to, įrodėme, kad bet kuri vienataškė aibė priklauso σ -algebrai $\mathcal{B}$, todėl ir šių dviejų aibių suma $[a; b]$ priklauso $\mathcal{B}$.

Teorema įrodyta.

2. Formulės (1) pagrindu pastebėsime, kad σ -algebra yra aibių klasė, uždara ne tik sumos, bet ir sankirtos operacijų atlikimo begalinį skaičių kartu atžvilgiu.

Jei Ω yra bet kokia aibė, o $\mathcal{F}$ – jos poaibių σ -algebra, tai dvejetas $(\Omega, \mathcal{F})$ vadinamas mačiaja erdve¹.

Jau pats šios erdvės pavadinimas sufleruoja, kad viskas paruošta tikimybinių mato sąvokos įvedimui.

3. Funkcija P , apibrėžta σ -algebroje $\mathcal{F}$ ir įgijanti reikšmes iš intervalo $[0; 1]$, vadinama **tikimybinio mato** mačiojoje erdvėje $(\Omega, \mathcal{F})$, jei:

$$(P1) \quad P(A) \geq 0 \quad \forall A \in \mathcal{F};$$

$$(P2) \quad P(\Omega) = 1;$$

$$(P3) \quad \forall A_i \in \mathcal{F} \text{ tokiems, kad } A_i \cap A_j = \emptyset \ (i \neq j),$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

4. **Pavyzdys.** Išnagrinėsime tokį stochastinį eksperimentą – „taškas“ metas į atkarpą $[0; 1)$. Tarkime, kad visi galimi rezultatai yra vienodai galimi. Aišku, kad šiame stochastiniame eksperimente $\Omega = [0; 1)$.

Tačiau jei, kaip dažnai darėme anksčiau, kiekvienam elementariam įvykiui ω mėginsime priskirti tikimybę $P(\omega)$, tai nieko gero negausime. Iš tiesų, kadangi visi elementarieji įvykiai „vienodai galimi“ ir jų yra be galo daug, tai $P(\omega) = 1/\infty = 0 \quad \forall \omega \in \Omega$.

Todėl tikimybinį matą P reikia konstruoti kitaip. Tikimybę „taškui“ patekti į intervalą $[a; b) \subset [0; 1)$ apibrėžkime lygią skaičiui $b - a$. Sumai nesikertančių intervalų $[a_i; b_i)$,

$$\bigcup_{i=1}^m [a_i; b_i) \subset [0; 1)$$

tokiu atveju priskirtina tikimybė

$$\sum_{i=1}^m (b_i - a_i).$$

Mačioji erdvė $(\Omega, \mathcal{F})$ pilnai aprašo nagrinėjamą stochastinį eksperimentą, jei $\mathcal{F}$ apibrėšime kaip Borelio σ -algebrą intervale $[0; 1)$.

¹ Angl. *measurable space*, rus. *измеримое пространство*.

28. KOLMOGORO AKSIOMŲ SISTEMA. TIKIMYBINĖ ERDVĖ

1. Pagal tarptautinių žodžių žodyną, *aksioma*¹ – tai pradinis išeities teiginys, sudarantis kitų teiginių (teoremų) įrodymų pagrindą mokslinėje teorijoje, kurios ribose priimamas be įrodymo.

Vaizdžiai kalbant, *aksiomų sistema* – tai nagrinėjamos teorijos fundamentas. Šiame skyrelyje išnagrinėsime statistikos (o tuo pačiu ir tikimybių teorijos) fundamentą, ant kurio ir yra statomas didžiulis statistikos rūmas.

Nagrinėjimui pasirinkime bet kokią stochastinį eksperimentą. Tegu Ω – šio stochastinio eksperimento *elementariųjų įvykių erdvė*. Tarkime, kad Ω poaibių sistema $\mathcal{F}$ yra σ -algebra. Tai reiškia, kad

(F1) $\Omega \in \mathcal{F}$;

(F2) Jei $A \in \mathcal{F}$, tai $\bar{A} \in \Omega \setminus A \in \mathcal{F}$;

(F2) Jei $A_i \in \mathcal{F}$, $i = 1, 2, \dots$, tai $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Aibės iš $\mathcal{F}$ vadinamos *atsitiktiniais įvykiais*, o dvejetas $(\Omega, \mathcal{F})$, kaip buvo minėta, *mačiąja erdve*.

2. Šioje erdveje $(\Omega, \mathcal{F})$ kiekvienam atsitiktiniam įvykiui A (t.y. aibei iš klasės $\mathcal{F}$) priskirsime skaičių $P(A)$, kuris vadinamas *atsitiktinio įvykio A tikimybe*, taip, kad būtų patenkintos praeito skyrelio sąlygos $(P1)$, $(P2)$ ir $(P3)$.

Apibrėžimas. Teiginiai $(F1)$, $(F2)$, $(F3)$ ir $(P1)$, $(P2)$, $(P3)$ vadinami *statistikos aksiomų sistema*.

Tokiu pavidalu šią sistemą pirmasis suformulavo rusų matematikas A. Kolmogorovas, todėl dažnai ji vadinama *Kolmogorovo aksiomų sistema*.

Aksimos $(P1)$ ir $(P3)$ reiškia, kad funkcija $P(\cdot)$, apibrėžta σ -algebroje $\mathcal{F}$, yra *matas*. Kadangi šis matas tenkina dar vieną sąlygą – sąlygą $(P2)$, – tai jis vadinamas *normuotu* (arba *tikimybiniu*) *matu*.

Apibrėžimas. Trejetas $(\Omega, \mathcal{F}, P)$, kuriame $\mathcal{F}$ pažymėta aibės Ω poaibių sistemos σ -algebra, o $P(\cdot)$ – tikimybinis matas σ -algebroje $\mathcal{F}$, yra vadinamas *tikimybine erdve*.

3. Išnagrinėsime paprasčiausias tikimybinio mato savybes.

Teorema 1. Tegu A ir B – atsitiktiniai įvykiai, t.y. $A \in \mathcal{F}$ ir $B \in \mathcal{F}$. Tegu, be to, $A \subset B$. Tada

$$P(B \setminus A) = P(B) - P(A). \quad (1)$$

¹ Gr. *axiōma*.

Įrodymas. Pagal tikimybinės erdvės apibrėžimą matas $P(\cdot)$ yra apibrėžtas tik σ -algebroje $\mathcal{F}$, todėl pirmiausia reikia įrodyti, kad $B \setminus A \in \mathcal{F}$. Tuo tikslu pastebėsime, kad jei $A \subset B$, tai $B \setminus A = \bar{A} \cap B$. Pagal aksiomą ($F2$) turime, kad $\bar{A} \in \mathcal{F}$, o pagal 26-ojo skyrelio teiginį 1 – dviejų σ -algebros $\mathcal{F}$ elementų (mūsų atveju tai aibės $\bar{A}$ ir B) sankirta vėl priklauso $\mathcal{F}$. Taigi,

$$B \setminus A = \bar{A} \cap B \in \mathcal{F}.$$

Kadangi $A \subset B$, tai

$$B = A \cup (B \setminus A), \quad A \cap (B \setminus A) = \emptyset,$$

Pagal aksiomą ($P3$) iš čia gauname, kad

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

Teorema įrodyta.

4. Pateiksime kai kurias šios teoremos išvadas.

Išvada 1. *Priešingojo įvykiui A įvykio $\bar{A}$ tikimybė yra $1 - P(A)$, t.y.*

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

Įrodymas akivaizdus – formulėje (1) pakanka paimti $B = \Omega$.

Išvada 2. *Negalimojo įvykio $\emptyset$ tikimybė yra lygi nuliui, t.y. $P(\emptyset) = 0$.*

Norint įsitikinti šios išvados teisingumu, formulėje (1) pakanka paimti $B = A$.

Išvada 3. *Jei $A \subset B$, tai $P(A) \leq P(B)$.*

Iš tiesų, pagal aksiomą ($P1$) $P(B \setminus A) \geq 0$. Pagal (1) iš čia gauname, kad

$$P(B) - P(A) \geq 0, \quad \text{t.y.} \quad P(A) \leq P(B).$$

Pastarojoje formulėje paėmę $B = \Omega$ iš aksiomos ($P2$) gauname tokį teiginį.

Išvada 4. *Jei A yra atsitiktinis įvykis, t.y. $A \in \mathcal{F}$, tai $P(A) \leq 1$.*

5. Pagal išvadą 2, $P(\emptyset) = 0$. Tačiau atvirkščias teiginys yra neteisingas, t.y. įvykis A , kurio tikimybė yra lygi nuliui, nebūtinai yra negalimasis įvykis. Norėdami tai pagrįsti, sukonstruosime konkretų pavyzdį.

Tegu $\Omega = [0; 1)$, o $\mathcal{F}$ – intervalų iš $[0; 1)$ Borelio σ -algebra. Intervalo $[a; b)$ (o taip pat bet kurio iš intervalų $(a; b)$, $[a; b]$, $(a; b]$) taip vadinamas *Lebego matas* m lygus šio intervalo ilgiui $b - a$. Apibrėžkime

$$P(A) = m(A).$$

Jei C yra atkarpos $[0; 1)$ racionalių skaičių aibė, tai

$$P(C) = m(C) = 0, \quad \text{tačiau} \quad C \neq \emptyset.$$

6. Skyrelio pabaigoje suformuluosime ir įrodysime taip vadinamą įvykių sumos teoremą.

Teorema 2. Tegu $A \in \mathcal{F}$ ir $B \in \mathcal{F}$. Tada

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

Įrodymas. Pagal aksiomą (F3), $A \cup B \in \mathcal{F}$. Taigi, yra prasmė kalbėti apie šios aibės matą $P(A \cup B)$.

Pastebėsime, kad aibę $A \cup B$ galima suskaidyti į trijų nesikertančių aibių sumą:

$$A \cup B = (A \cap B) \cup (A \setminus (A \cap B)) \cup (B \setminus (A \cap B)).$$

Todėl iš aksiomos (P3) ir teoremos 1 gauname, kad

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + P(A \setminus (A \cap B)) + \\ &\quad + P(B \setminus (A \cap B)) = P(A \cap B) + P(A) - P(A \cap B) + \\ &\quad + P(B) - P(A \cap B) = P(A) + P(B) - P(A \cap B). \end{aligned}$$

Teorema įrodyta.

29. ATSITIKTINIS DYDIS KAIP MAČIOJI FUNKCIJA IR JO GENERUOTI ATSITIKTINIAI ĮVYKIAI

1. Paprasčiausiu atveju, kaip mes jau matėme 5-ajame skyrelyje, *atsitiktinis dydis* – tai funkcija, suteikianti kiekvienai stochastinio eksperimento baigčiai skaitinę reikšmę. Tai reiškia, kad atsitiktinį dydį X galima nagrinėti kaip įgyjančią realias reikšmes funkciją $X = X(\omega)$, apibrėžtą elementariųjų įvykių erdvėje Ω .

Tačiau norint matematiškai griežtai formuluoti statistikos teiginius ir teoremas, tokio apibrėžimo nepakanka.

2. Tai galima pagrįsti tuo, kad tikimybių teorijoje yra parodoma, jog atsitiktinis dydis $X(\omega)$ pilnai nusakomas aibėmis $\{\omega \mid X(\omega) < x\}$, kur $x \in \mathbb{R}^1$. Šios aibės tikimybiniėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$ yra „informatyvios“ tik tuo atveju, jei jas galima išmatuoti. O išmatuoti jas šioje erdvėje galima tik tada, kai bet kuriam realiam x

$$\{\omega \mid X(\omega) < x\} \in \mathcal{F}, \quad (1)$$

nes pagal tikimybės erdvės apibrėžimą matas $\mathbf{P}$ yra apibrėžtas σ -algebroje $\mathcal{F}$.

3. Suformuluosime pagrindinį šio skyrelio apibrėžimą.

Realioji funkcija $X = X(\omega)$, apibrėžta aibėje Ω , vadinama atsitiktiniu dydžiu tikimybiniėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, jei bet kuriam realiam x tenkinamas sąryšis (1).

Funkcijos, tenkinančios sąryšį (1), funkcijų teorijoje vadinamos *mačiosiomis atžvilgiu σ -algebros $\mathcal{F}$ funkcijomis*.

Pateiksime tokios funkcijos pavyzdį.

4. **Pavyzdys.** Stochastinį eksperimentą apibrėžkime tokiu būdu: moneta metama vieną kartą. Tada $\Omega = \{\omega_1, \omega_2\}$, kur ω_1 – herbo pasirodymo elementarus įvykis, o ω_2 – skaičiaus pasirodymo elementarusis įvykis. Tarkime, kad $\mathcal{F}$ – visų galimų Ω poaibių aibė ir $\mathbf{P}(\{\omega_1\}) = \mathbf{P}(\{\omega_2\}) = 1/2$. Funkciją $X(\omega)$ apibrėžkime taip:

$$X(\omega) = \begin{cases} 1, & \text{kai } \omega = \omega_1, \\ 0, & \text{kai } \omega = \omega_2. \end{cases}$$

Pastebėsime, kad

$$\{\omega \mid X(\omega) < x\} = \begin{cases} \emptyset, & \text{kai } x \leq 0, \\ \{\omega_2\}, & \text{kai } 0 < x \leq 1, \\ \Omega, & \text{kai } 1 < x. \end{cases}$$

Tai reiškia, kad $\{\omega \mid X(\omega) < x\} \in \mathcal{F}$ bet kuriam realiam x , t.y. kad $X(\omega)$ yra atsitiktinis dydis nagrinėjamojoje tikimybiniėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$.

5. Nagrinėjant atsitiktinius dydžius, dažnai tenka atsakyti į tokį klausimą: kokia yra tikimybė įvykio, kad atsitiktinio dydžio $X(\omega)$ reikšmės priklauso bet kuriai Borelio aibei B ?

Tai reiškia, kad pakankamai plačiai aibių klasei $\{B\}$ mes turime patikrinti, kad

$$\{\omega \mid X(\omega) \in B\} \in \mathcal{F}, \quad (2)$$

nes tik tokiu atveju yra apibrėžtas matas P ir yra prasmė nagrinėti tikimybę $P(\{\omega | X(\omega) \in B\})$.

Apibrėžimas. Jei $X(\omega)$ yra atsitiktinis dydis tikimybiniame erdvėje $(\Omega, \mathcal{F}, P)$, o B – bet kokia Borelio aibė, tai aibė $\{\omega | X(\omega) \in B\}$ vadinama atsitiktiniu įvykiu, kurį generavo atsitiktinis dydis $X(\omega)$.

Perskaičius šį apibrėžimą, iškyla natūralus klausimas – ar aibė $\{\omega | X(\omega) \in B\}$ tikrai yra atsitiktinis įvykis, t.y. ar tikrai

$$X^{-1}(B) = \{\omega | X(\omega) \in B\} \in \mathcal{F}?$$

6. Atsakymas į šį klausimą yra teigiamas, tačiau jo įrodymą pateiksime tik pagrindiniam atvejui, t.y. kai B yra arba intervalas, arba pusašė, arba taškas.

Teorema. Tegu $X(\omega)$ – atsitiktinis dydis tikimybiniame erdvėje $(\Omega, \mathcal{F}, P)$. Tada kiekvienas iš sekančių Ω poaibių

$$\begin{aligned} &\{\omega | X(\omega) \geq x\}, \quad \{\omega | X(\omega) \leq x\}, \quad \{\omega | X(\omega) > x\}, \\ &\{\omega | x \leq X(\omega) < y\}, \quad \{\omega | x < X(\omega) < y\}, \quad \{\omega | X(\omega) = x\} \end{aligned}$$

bet kuriems realiems skaičiams x ir y yra atsitiktinis įvykis, t.y. kiekviena iš šių aibių priklauso σ -algebrai $\mathcal{F}$.

Įrodymas grindžiamas, pirma, atsitiktinio įvykio apibrėžimu ir, antra, σ -algebros apibrėžimu bei jos savybėmis. Tai, kad $X(\omega)$ yra atsitiktinis dydis, reiškia, jog bet kuriam realiam x galioja sąryšis (1). Teoremoje išvardytus Ω poaibių stengsimės pavaizduoti kaip aibių tipo (1) sumą, neigimą, sankirtą arba skirtumą, nes σ -algebra yra uždara šių operacijų atžvilgiu, taigi, $\mathcal{F}$ jų atžvilgiu taip pat uždara. Šiam tikslui pakanka pastebėti, kad

$$\begin{aligned} &\{\omega | X(\omega) \geq x\} = \Omega \setminus \{\omega | X(\omega) < x\}, \\ &\{\omega | X(\omega) \leq x\} = \bigcap_{n=1}^{\infty} \{\omega | X(\omega) < x + 1/n\}, \\ &\{\omega | X(\omega) > x\} = \Omega \setminus \{\omega | X(\omega) \leq x\}, \\ &\{\omega | x \leq X(\omega) < y\} = \{\omega | X(\omega) < y\} \setminus \{\omega | X(\omega) < x\}, \\ &\{\omega | x < X(\omega) < y\} = \{\omega | X(\omega) < y\} \setminus \{\omega | X(\omega) \leq x\}, \\ &\{\omega | X(\omega) = x\} = \{\omega | X(\omega) \leq x\} \setminus \{\omega | X(\omega) < x\}. \end{aligned}$$

Teorema įrodyta.

30. TIKIMYBIŲ SKIRSTINYS. SKIRSTINIO FUNKCIJA IR JOS SAVYBĖS

1. Praeitame skyrelyje matėme, kad jei $X(\omega)$ yra atsitiktinis dydis tikimybėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, tai

$$\{\omega | X(\omega) \in B\} \in \mathcal{F}$$

bet kuriai Borelio aibei B . Tai reiškia, kad bet kuriai Borelio aibei B egzistuoja matas

$$\mathbf{P}(\{\omega | X(\omega) \in B\}), \quad (1)$$

kuris yra vadinamas atsitiktinio dydžio X **tikimybių skirstiniu**¹.

Kadangi B , kaip minėta, yra bet kokia Borelio aibė, tai išraišką (1) galima nagrinėti kaip aibių funkciją. Lietuvos standarte [8] pateikiamas toks apibrėžimas.

Tikimybių skirstinys – tai funkcija, kuri duoda tikimybę, kad atsitiktinis dydis įgyja nurodytąją reikšmę arba priklauso nurodytajai reikšmių aibei.

Suprantama, toks apibrėžimas nėra pakankamai griežtas matematine prasme, bet naudingas terminų pasirinkimo kontekste.

Praeito skyrelio teoremoje matėme, kad aibė $\{\omega | X(\omega) = x\}$ priklauso σ -algebrai $\mathcal{F}$. Būtent ši aibė yra tai, kas standarto apibrėžime išreiškiama žodžiais „atsitiktinis dydis įgyja nurodytąją reikšmę“.

2. Taigi, tikimybių skirstiniai yra Borelio aibių funkcijos. Todėl juos sunku nagrinėti klasikinės analizės metodais, nes ši analizė išvystyta „taškų“ (t.y. vieno kintamojo) funkcijoms. Siekiant plačiai taikyti klasikinius analizinius metodus statistikoje, tikimybių teorijoje įrodomas abipus vienareikšmio sąryšio tarp tikimybinių skirstinių (1) ir taip vadinamų *skirstinio funkcijų* $\mathbf{P}(X < x)$, kur $x \in (-\infty; \infty)$, egzistavimas.

Skirstinio funkcija² – tai funkcija, kuri kiekvienai x reikšmei priskiria tikimybę, kad atsitiktinis dydis X bus ne didesnis už x :

$$F(x) = \mathbf{P}(X < x).$$

Tai Lietuvos standarto apibrėžimas. Vietoj jame naudojamo termino *skirstinio funkcija* mes vartojame sutrumpinimą **skirstinys**.

¹ Angl. *probability distribution*, rus. *распределение вероятностей*.

² Angl. *distribution function*, rus. *функция распределения*.

3. Panagrinėsime skirstinių pavyzdžius.

Pavyzdys 1. Stochastinis eksperimentas yra toks: į intervalą $[a; b]$ atsitiktinai metamas taškas. Skaitoma, kad bet kuri taško padėtis šiame intervale yra vienodai galima.

Šiuo atveju $\Omega = [a; b]$, $\mathcal{F}$ yra pusiau atvirų iš dešinės intervalų iš $[a; b]$ Borelio σ -algebra, o tikimybinių matą $\mathbf{P}$ galima įvesti tokiu būdu:

$$\mathbf{P}([\alpha; \beta]) = \frac{\beta - \alpha}{b - a}, \quad \text{jei } [\alpha; \beta] \subset [a; b].$$

Funkciją $X = X(\omega)$ apibrėšime taip: $X(\omega) = \omega$, jei $\omega \in [a; b]$. Tada

$$\{\omega | X(\omega) < x\} = \begin{cases} \emptyset, & \text{jei } x \leq a, \\ [a; x], & \text{jei } a < x \leq b, \\ \Omega = [a; b], & \text{jei } x > b. \end{cases}$$

Iš čia darome išvadą, kad $\{\omega | X(\omega) < x\} \in \mathcal{F}$ bet kuriam realiam x , t.y. funkcija $X(\omega)$ yra mačioji funkcija tikimybiniame erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, o tuo pačiu ji yra ir atsitiktinis dydis šioje erdvėje.

Nesunku įsitikinti, kad šio atsitiktinio dydžio skirstinys yra taip vadinamas *tolygusis skirstinys* (kurio tankį nagrinėjome 9-ajame skyrelyje):

$$F(x) = \mathbf{P}(X < x) = \begin{cases} 0, & \text{jei } x \leq a, \\ \frac{x - a}{b - a}, & \text{jei } a < x \leq b, \\ 1, & \text{jei } x > b. \end{cases}$$

Šio skirstinio grafikas pateikiamas 42 pav.

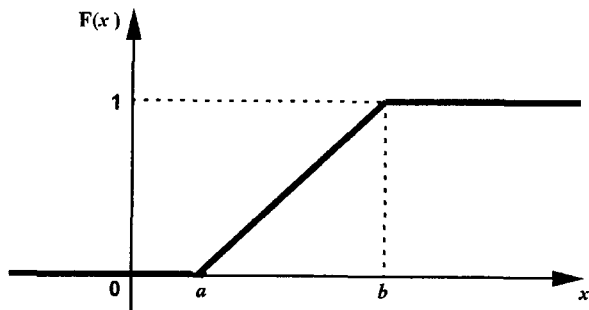
Pavyzdys 2. Apskaičiuosime atsitiktinio dydžio X , apibrėžto 29-ojo skyrelio 4-ajame punkte, skirstinį. Kadangi šiuo atveju

$$\mathbf{P}(\{\omega_1\}) = \mathbf{P}(\{\omega_2\}) = 1/2$$

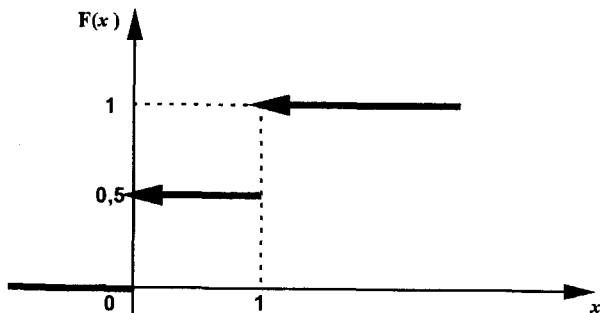
ir visada $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(\Omega) = 1$, tai

$$F(x) = \begin{cases} 0, & \text{kai } x \leq 0, \\ \frac{1}{2}, & \text{kai } 0 < x \leq 1, \\ 1, & \text{kai } 1 < x. \end{cases}$$

Ši funkcija $F(x)$ vadinama *dvitaškiu skirstiniu*, o jo grafikas pateikiamas 43 pav.



42 pav. Tolygiojo intervale $[a; b]$ skirstinio grafikas



43 pav. Dvitaškio skirstinio grafikas

4. Ar skirstinio apibrėžimas yra korektiškas, t.y. ar tikimybė $P(X < x)$ visada egzistuoja?

Atsakymas teigiamas tuo atveju, jei aibė $\{\omega | X(\omega) < x\}$ yra atsitiktinis įvykis, t.y. jei ji priklauso σ -algebrai $\mathcal{F}$. O taip bus tada, kai X yra atsitiktinis dydis tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, t.y. kai X yra mačioji funkcija atžvilgiu σ -algebros $\mathcal{F}$.

5. Panagrinėsime paprasčiausias atsitiktinio dydžio $X = X(\omega)$ tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$ skirstinio $F(x) = \mathbf{P}(\{\omega | X(\omega) < x\})$ savybes.

1 savybė. Jei $a < b$, tai

$$\mathbf{P}(\{\omega : a \leq X(\omega) < b\}) = F(b) - F(a). \quad (1)$$

2 savybė. $F(x)$ yra nemažėjanti funkcija, t.y. jei $a < b$, tai $F(a) \leq F(b)$.

3 savybė. $\lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0$, $\lim_{x \rightarrow +\infty} F(x) = F(+\infty) = 1$.

4 savybė. $F(x)$ yra tolydi iš kairės.

Pirmąsias dvi savybes įrodysime.

Kadangi

$$\{\omega | a \leq X(\omega) < b\} = \{\omega | X(\omega) < b\} \setminus \{\omega | X(\omega) < a\}$$

ir, be to, iš sąlygos $a < b$ išplaukia, kad

$$\{\omega | X(\omega) < a\} \subset \{\omega | X(\omega) < b\},$$

tai pagal tikimybinio mato savybę (žr. teoremą 1 28-ajame skyrelyje) gauname formulę (1).

2 savybė įrodoma 1 savybės pagrindu:

$$F(b) - F(a) = \mathbf{P}(\{\omega | a \leq X(\omega) < b\}) \geq 0.$$

6. Pasirodo, kad yra teisingas atvirkščias teiginys. Jei bet kuri funkcija tenkina išvardytas savybes (pakanka 2, 3 ir 4 savybių), tai ji yra skirstinio funkcija. Suformuluosime tai be įrodymo.

Teorema. Tegu funkcija $F(x)$ tenkina tokias tris sąlygas: 1) $F(x)$ yra nemažėjanti tiesėje $(-\infty, \infty)$; 2) $F(x)$ yra tolydi iš kairės; 3) $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow +\infty} F(x) = 1$. Tada egzistuoja tokia tikimybinė erdvė $(\Omega, \mathcal{F}, \mathbf{P})$ ir joje apibrėžtas atsitiktinis dydis $X(\omega)$, kad $X(\omega)$ skirstinys yra $F(x)$.

7. Pirmojoje vadovėlio dalyje daug dėmesio skyrėme tankiams. Pateiksime matematiškai griežtą šios sąvokos apibrėžimą.

Sakoma, kad atsitiktinis dydis $X(\omega)$ su skirstiniu $F(x)$ turi tankį, jei egzistuoja tokia integruojama funkcija $p(u)$, kad visiems realiems x

$$F(x) = \int_{-\infty}^x p(u) \, du.$$

Funkcija $p(u)$ vadinama atsitiktinio dydžio $X(\omega)$ tankiu.

Pabaigai pastebėsime, kad jei $p(u)$ – tolydi, tai iš analizės kurso seka, jog

$$p(x) = \frac{dF(x)}{dx}.$$

31. VEIKSMAS BE PRAEITIES POVEIKIO. KOŠI LYGTIS STATISTIKOJE

1. Tegu X yra bet koks neneigiamas atsitiktinis dydis, kurį galima interpretuoti kaip prietaiso gyvavimo laiką (pavyzdžiui, elektros lemputės tarnavimo laikotarpį) arba kaip aptarnavimo trukmę (pavyzdžiui, telefoninio pokalbio trukmės laiką).

Tada $P(X \geq x)$ yra tikimybė įvykio, kad „prietaiso gyvavimo laikas yra ne mažesnis už x “. Tarkime, kad prietaisas dirbo laiko tarpą y , ir nuo jo, šito laiko tarpo y , likęs prietaiso gyvavimo amžius (skaičiuojant nuo laiko momento $x + y$) nepriklauso, t.y.

$$P(X \geq x + y | X \geq y) = P(X \geq x). \quad (1)$$

Ši savybė anglų kalba vaizdžiai vadinama *lack of memory property* (pažodžiui verčiant – *atminties neturėjimo savybė*), o rusų kalba – *свойство отсутствия последствств* (**veiksmo be praeities poveikio savybė**). Mes pasirinksimė pastarąjį terminą.

2. Pasinaudojus sąlyginės tikimybės apibrėžimu lengva įsitikinti, kad eksponentinis skirstinys, apibrėžiamas formule

$$P(X < x) = \begin{cases} 1 - \exp(-\lambda x), & \text{jei } x \geq 0, \\ 0, & \text{jei } x < 0, \end{cases}$$

($\lambda > 0$ yra parametras), tenkina sąryšį (1). Ar ši eksponentinių skirstinių klasė yra viernintelė klasė, kurios elementai turi šią savybę? Tai pagrindinis šio skyrelio klausimas, o teoremos, kuriose aprašoma (charakterizuojama)

skirstinių klasė nagrinėjamų atsitiktinių dydžių funkcijų savybėmis vadinamos *charakterizacijos teoremomis*.

Perrašysime sąryšį (1) kita, mums patogesne forma.

Jei $P(X = 0) \neq 1$ (t.y. jei atsitiktinis dydis X yra neišsigimęs nulyje), tai egzistuoja toks skaičius $y_0 > 0$, kad $P(X \geq y_0) > 0$. Įrodysime, kad ir

$$P(X \geq ny_0) > 0, \quad n = 1, 2, \dots \quad (2)$$

3. Įrodinėsime matematinės indukcijos metodu. Jau matėme, kad kai $n = 1$, tai (2) yra teisinga. Tegu ji yra teisinga tada, kai $n = k$. Tada iš (1), paėmę $x = ky_0$ ir $y = y_0$ gauname, kad

$$P(X \geq (k+1)y_0) = P(X \geq ky_0)P(X \geq y_0) > 0.$$

Taigi, formulė (2) yra teisinga ir reikšmei $n = k+1$, o tuo pačiu kiekvienam natūriniam n . O kadangi bet kuriam $y > 0$ egzistuoja n toks, kad $P(X \geq y) \geq P(X \geq ny_0)$, tai $P(X \geq y) > 0$.

Todėl prisiminę sąlyginės tikimybės apibrėžimą, veiksmo be praeities poveikio savybę (1) galime perrašyti taip:

$$P(X \geq x+y) = P(X \geq x)P(X \geq y), \quad \forall x \geq 0, \forall y \geq 0. \quad (3)$$

4. Analizės kurse funkcionalinė lygtis

$$f(x+y) = f(x)f(y), \quad \forall (x, y) \in S \subseteq \mathbb{R}^2 \quad (4)$$

yra vadinama *Koši lygtimi*. Paėmę $S = [0; \infty) \times [0; \infty)$ gauname, kad **veiksmo be praeities poveikio savybė (1) yra ekvivalenti Koši lygčiai (4)**, jei $f(x) = P(X \geq x)$.

Pasirodo, kad eksponentinio dėsnio charakterizacijai pakanka veiksmo be praeities poveikio savybės (3) išpildymo ne visuose taškuose $y \geq 0$, o tik dviejuose nebendramačiuose taškuose y_1 ir y_2 . Realieji skaičiai y_1 ir y_2 vadinami *nebendramačiais*, jei jų santykis y_1/y_2 yra iracionalus skaičius.

5. **Charakterizacijos teorema¹**. Jeigu funkcija $f(x)$ nėra konstanta ir bent dviejuose nebendramačiuose taškuose $0 < y_1 < y_2$ išpildyti sąryšiai $f(y_i) > 0$ ir šiuose taškuose tenkinama Koši lygtis, t.y.

$$f(x+y_i) = f(x)f(y_i), \quad \forall x \geq 0, i = 1, 2, \quad (5)$$

¹ Medžiaga dėstoma pagal O. Januškevičienės straipsnį (žr. [11], psl. 220).

tai egzistuoja $\lambda > 0$ toks, kad $f(x) = \exp(-\lambda x) \forall x \geq 0$.

Irodymas. Įvedę naujus žymėjimus

$$\lambda_i = -\frac{1}{y_i} \ln f(y_i), \quad \varphi_i(x) = \exp(\lambda_i x) f(x),$$

iš sąlygos (5) gauname, kad $\forall x \geq 0, i = 1, 2$

$$\varphi_i(x + y_i) \exp(-\lambda_i x - \lambda_i y_i) = \varphi_i(x) \exp(-\lambda_i x) \varphi_i(y_i) \exp(-\lambda_i y_i).$$

Tai reiškia, kad $\forall x \geq 0$ ir $i = 1, 2$

$$\varphi_i(x + y_i) = \varphi_i(x) \varphi_i(y_i). \quad (6)$$

Pasinaudoję λ_i apibrėžimu pastebime, kad

$$\varphi_i(y_i) = \exp(\lambda_i y_i) f(y_i) = \exp \left\{ -\frac{y_i}{y_i} \ln f(y_i) \right\} f(y_i) = 1.$$

Iš čia ir iš (6) gauname, kad

$$\varphi_i(x + y_i) = \varphi_i(x), \quad i = 1, 2, \quad \forall x \geq 0. \quad (7)$$

Taigi, pusašėje $[0; \infty)$ $\varphi_1(x)$ yra periodinė funkcija su periodu y_1 , o $\varphi_2(x)$ – su periodu y_2 .

Irodysime dabar, kad $\lambda_1 = \lambda_2$. Tarkime, kad taip nėra, t.y. tegu $\lambda_1 \neq \lambda_2$. Kadangi iš $\varphi_1(x)$ ir $\varphi_2(x)$ apibrėžimo seka, kad

$$f(x) = \exp(-\lambda_1 x) \varphi_1(x) = \exp(-\lambda_2 x) \varphi_2(x) \quad \forall x \geq 0,$$

tai visiems $x \geq 0$

$$\varphi_1(x) = \exp((\lambda_1 - \lambda_2)x) \varphi_2(x).$$

Tačiau tada prielaida $\lambda_1 \neq \lambda_2$ prieštariauja tam, kad $\varphi_1(x)$ yra periodinė funkcija.

Taigi, $\lambda_1 = \lambda_2$ ir $\varphi_1(x) = \varphi_2(x) = \varphi(x) \forall x \geq 0$.

Sąryšis (7) reiškia, kad funkcija $\varphi(x)$ turi nebendramačius periodus, o tai įmanoma tik tuo atveju, kai $\varphi(x) \equiv c \forall x \geq 0$, kur c yra konstanta.

Iš (5) pastebime, kad $f(0) = 1$. Tada pagal $\varphi_i(x)$ apibrėžimą turime, kad ir $\varphi_i(0) = 1$, t.y. ir $\varphi(0) = 1$. Tai reiškia, kad $c = 1$. Dabar belieka vėl pasinaudoti $\varphi_i(x)$ apibrėžimu, iš kur gauname, kad

$$f(x) = c \exp(-\lambda x) = \exp(-\lambda x), \quad \forall x \geq 0;$$

čia $\lambda = \lambda_1 = \lambda_2$.

Teorema įrodyta.

6. Skyrelio pabaigoje pastebėsime, kad eksponentinio dėsnio tankis yra toks:

$$p(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{jei } x \geq 0, \\ 0, & \text{jei } x < 0. \end{cases}$$

32. EMPIRINIS MATAS. TEOREMA APIE ŠIŲ MATŲ SEKOS KONVERGAVIMĄ PAGAL TIKIMYBĘ

1. Pirmojoje vadovėlio dalyje matėme, koks svarbus statistikoje yra *skirstinių parametrų vertinimo uždavinys*. Mūsų kelių artimiausių skyrelių tikslas – pagal galimybes matematiškai griežtai išdėstyti šio uždavinio sprendimo matematinis metodus.

2. Daroma prielaida, kad nežinomas skirstinys priklauso skirstinių šeimai $\{f(x, \theta) \mid \theta \in \Theta\}$, kur θ yra parametras, o Θ – aibė tiesėje arba erdvėje $\mathbb{R}^n$. Tai reiškia, kad skirstinys priklauso nuo vieno arba kelių realių parametrų. Pavyzdžiui, normalių skirstinių šeima priklauso nuo dviejų parametrų – vidurkio μ ir dispersijos σ^2 .

Parametro θ vertinimo uždavinio esmė – turint imtį $X_1, \dots, X_n$ sukonstruoti apytikres formules

$$\theta \approx h(X_1, \dots, X_n), \quad (1)$$

kur $h(x)$ – mačioji (Borelio prasme) funkcija, apibrėžta erdvėje $\mathbb{R}^n$ su reikšmėmis aibėje Θ .

3. Apibrėžimas. Bet kurią imties $X_1, \dots, X_n$ mačiąją funkciją $h(X_1, \dots, X_n)$ vadiname statistika. Jei, be to, ši funkcija kažkuria prasme tenkina sąryšį (1), tai $h(X_1, \dots, X_n)$ vadinama parametro θ įvertiniu.

Čia iškart iškyla tokie klausimai. Kaip vertinti sąryšio (1) tikslumą? Kokius įvertinius h skaityti „gerais“? Kaip konstruoti konkrečius „geriausius“ įvertinius?

Atsakymų į šiuos klausimus mes ieškosime kituose skyreliuose, o dabar pereisime prie empirinio mato nagrinėjimo.

4. Tegū $X_1, \dots, X_n$ yra imtis, t.y. vienas ar daugiau ėmimo vienetų, paimtų iš populiacijos, kurios tikimybių skirstinys yra P . Šią imtį, kurios dydis

yra n , nagrinėsime kaip vektorių $\mathbf{X}_n$,

$$\mathbf{X}_n = (X_1, \dots, X_n),$$

kur $X_1, \dots, X_n$ yra nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai.

Tegu dabar B – Borelio aibė tiesėje. Apibrėžkime

$$\mathbf{P}_n^*(B) = \nu(B)/n, \quad (2)$$

kur $\nu(B)$ yra imties $\mathbf{X}_n$ (iš populiacijos su tikimybių skirstiniu P) elementų, papuolusių į aibę B , skaičius.

Dydis $\mathbf{P}_n^*(B) = \nu(B)/n$ yra vadinamas empiriniu matu, sukonstruotu imties $\mathbf{X}_n = (X_1, \dots, X_n)$ pagrindu.

5. Vietoj formulės (2) galima kita užrašymo forma. Tegu $I_X(B)$ – atsitiktinio dydžio X indikatorius, t.y.

$$I_X(B) = \begin{cases} 1, & \text{kai } X \in B, \\ 0, & \text{kai } X \notin B. \end{cases}$$

Galima įrodyti, kad $I_X(B)$ – atsitiktinis dydis ir, be to,

$$\nu(B) = \sum_{i=1}^n I_{X_i}(B)$$

Tada (2) galima perrašyti taip:

$$\mathbf{P}_n^*(B) = \frac{1}{n} \sum_{i=1}^n I_{X_i}(B).$$

Nagrinėdami tikimybinių mato įvedimą, 4-ajame skyrelyje pastebėjome, kad:

Atsitiktinio įvykio A santykinis dažnis $n(A)/n$ stochastinių eksperimentų serijoje turi tendenciją, esant pakankamai dideliame n , artėti prie pastovios reikšmės, kurią pažymėsime $\mathbf{P}(A)$. Skaičius $\mathbf{P}(A)$ vadinamas atsitiktinio įvykio A tikimybe.

Iškyla natūralus klausimas – ar turi tokią tendenciją santykis $\nu(B)/n$, kurį taip pat galima vadinti *santykiniu dažniu*? Ir jei atsakymas teigiamas, tai būtina kažkuria prasme šią tendenciją „išmatuoti“. Šių klausimų nagrinėjimu ir užsiimsime kitame punkte.

6. *Praktiškai neabejotina, kad didelio skaičiaus nepriklausomų populiacijos stebinių aritmetiniai vidurkiai artėja prie šios populiacijos vidurkio μ . Taip buvo formuluojamas didžiųjų skaičių dėsnis 6-ajame skyrelyje. Tikimybių teorijos kurse sąvokos „praktiškai neabejotina“ ir „artėja“ paprastai yra konkretizuojamos, o didžiųjų skaičių dėsnis formuluojamas tokiu būdu:*

Jei $X_1, X_2, \dots, X_n, \dots$ yra nepriklausomi vienodai pasiskirstę atsitiktiniai dydžiai, o $DX_i \leq c$ (c – konstanta), tai bet kuriam $\varepsilon > 0$

$$P\left(\left\{\omega \mid \left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \varepsilon\right\}\right) \rightarrow 0, \quad (4)$$

kai $n \rightarrow \infty$. Čia $\mu = EX_1 = \dots = EX_n = \dots$

Formulėje (4) vietoj X_i nagrinėsime atsitiktinius dydžius $I_{X_i}(B)$. Pastebėsime, kad $I_{X_i}(B)$ yra dvitaškis atsitiktinis dydis, įgyjantis reikšmę 1 su tikimybe $P(X_i \in B)$ ir reikšmę 0 su tikimybe $P(X_i \notin B)$. Todėl

$$EI_{X_i}(B) = 1 \cdot P(X_i \in B) + 0 \cdot P(X_i \notin B) = P(X_i \in B),$$

$$DI_{X_i}(B) = EI_{X_i}^2(B) - (EI_{X_i}(B))^2 = 1 \cdot P(X_i \in B) - P^2(X_i \in B) \leq 1.$$

O kadangi $\mu = EI_{X_i}(B) = P(X_i \in B)$, tai iš (4) gauname, kad $\forall \varepsilon > 0$

$$P\left(\left\{\omega \mid \left|\frac{1}{n} \sum_{i=1}^n I_{X_i}(B) - P(X_i \in B)\right| > \varepsilon\right\}\right) \rightarrow 0,$$

kai $n \rightarrow \infty$. Dabar pasinaudoję (3) turime, kad $\forall \varepsilon > 0$

$$P(\{\omega \mid |P_n^*(B) - P(X_i \in B)| > \varepsilon\}) \rightarrow 0,$$

kai $n \rightarrow \infty$.

7. Tuo pačiu įrodėme tokią teoremą apie empirinių matų konvergavimą.

Teorema. *Tegu B – bet kokia Borelio aibė, o $X_1, \dots, X_n$ yra imtis iš populiacijos, kurios tikimybių skirstinys yra P . Tada bet kuriam $\varepsilon > 0$*

$$P(\{\omega \mid |P_n^*(B) - P(B)| > \varepsilon\}) \rightarrow 0,$$

kai $n \rightarrow \infty$, kur $P(B) = P(X_1 \in B) = \dots = P(X_n \in B) = \dots$

Šią teoremą galima formuluoti trumpiau, įvedus konvergavimo pagal tikimybę sąvoką.

Apibrėžimas. Sakoma, kad atsitiktinių dydžių seka $\{X_n\}$ konverguoja pagal tikimybę į atsitiktinį dydį X , jei bet kuriam $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbf{P}(\{|X_n - X| > \varepsilon\}) = 0.$$

Naudojant konvergavimo pagal tikimybę sąvoką, teoremą galima suformuluoti taip:

Bet kuriai Borelio aibei B

$$\mathbf{P}_n^*(B) \xrightarrow[\text{tikimybę}]{\text{pagal}} P(B).$$

33. VARIACINĖ EILUTĖ, EMPIRINIAI SKIRSTINIAI.

GLIVENKO-KANTELI TEOREMA

1. Tegu $\mathbf{P}_n^*$ yra imties $\mathbf{X}_n = (X_1, \dots, X_n)$ empirinis matas. Šios imties empiriniu skirstiniu vadinama funkcija

$$F_n^*(x) = \mathbf{P}_n^*((-\infty; x)).$$

Iš praėito skyrelio (2) formulės nesunku pastebėti, kad dydis $nF_n^*(x)$ – tai imties $\mathbf{X}_n$ elementų, mažesnių už x , skaičius. Šios pastabos pagrindu praktikoje F_n^* konstravimui naudojama tokia pastaba.

Imties $\bar{X}_n = (X_1, \dots, X_n)$ stebiniai $(x_1, \dots, x_n)$ rūšiuojami pagal didėjimą, t.y. sudaroma seka

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

kuri yra vadinama *variacione eilute*. Tada

$$F_n^*(x) = \frac{k}{n}, \quad \text{kai } x \in (x_{(k)}, x_{(k+1)}],$$

kur $k = 0, 1, \dots, n$, $x_{(0)} = -\infty$, $x_{(n+1)} = +\infty$. Be to, $F_n^*(x) = 0$, kai $x \leq x_{(1)}$ ir $F_{n+1}^*(x) = 1$, kai $x > x_{(n+1)}$.

Pastebėsime, kad $F_n^*(x)$ yra laiptuota funkcija su šuoliukais, kurių dydis yra $1/n$, taškuose x_i , jei visi x_i yra skirtingi.

16 lentelė

Partijos numeris	1	2	3	4	5	6	7	8	9	10
Nukrypimas nuo vidutinės reikšmės	-1	0	1	1	-1	1	0	-2	2	1

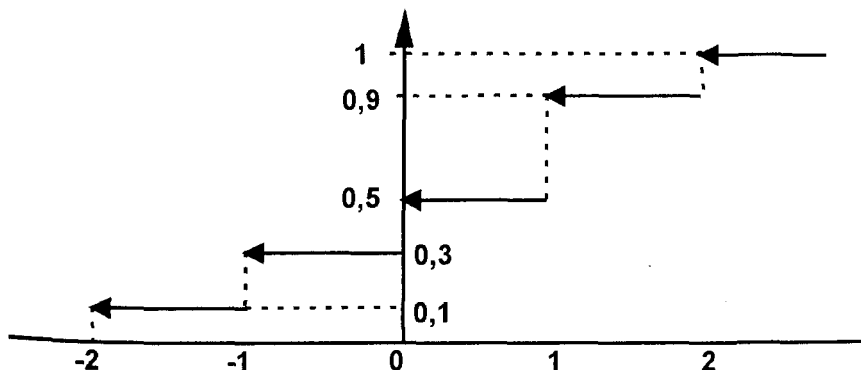
2. **Pavyzdys.** Kiekvienam šimtui televizorių, gaminamų koncerne, vidutiniškai du televizoriai neatitinka šiuolaikinių standartų. Atliktame dešimties partijų po šimtą televizorių patikrinime nustatyti tokie nukrypimai rastų brokuotų televizorių skaičiaus nuo vidutinės reikšmės:

Rasti šios imties variacinę eilutę ir nubrėžti jos empirinį skirstinį.

Sprendimas. Variacinė eilutė yra tokia:

$$-2, -1, -1, 0, 0, 1, 1, 1, 1, 2.$$

Skirtingos šios variacinės eilutės reikšmės yra šios: $-2, -1, 0, 1, 2$. Šių reikšmių kartotinumai atitinkamai yra 1, 2, 2, 4, 1. Tai reiškia, kad taške -2 empirinis skirstinys turės šuoliuką, kurio dydis yra $1/10$, taške -1 – dydžio $2/10$ šuoliuką, taške 0 – tokį pat šuoliuką. Didžiausias šuoliukas, kurio dydis $4/10$, yra taške 1 . Pagaliau, taške 2 yra dydžio $1/10$ šuoliukas. Dabar belieka nubrėžti grafiką:



44 pav. Imties, kurios dydis $n = 10$, empirinio skirstinio grafiko pavyzdys

3. Toliau mums prireiks konvergavimo su tikimybe 1 sąvokos.

Apibrėžimas. Atsitiktinių dydžių seka $\{X_n\}$ konverguoja su tikimybe 1 į atsitiktinį dydį X , jei

$$P\left(\left\{\omega \mid \lim_{n \rightarrow \infty} |X_n(\omega) - X(\omega)| = 0\right\}\right) = 1.$$

Yra žinoma, kad seka $X_n \rightarrow X$ su tikimybe 1 tada ir tik tada, kai kiekvienam $\varepsilon > 0$

$$P\left(\bigcup_{m=1}^{\infty} \{|X_{n+m} - X| \geq \varepsilon\}\right) \rightarrow 0, \quad \text{kai } n \rightarrow \infty.$$

Šio teiginio pagrindu įrodoma, kad iš sekos konvergavimo su tikimybe 1 išplaukia šios sekos konvergavimas pagal tikimybę. Atvirkščias teiginys, apskritai paėmus, yra neteisingas.

4. Sustiprinto didžiųjų skaičių dėsnio teoremomis vadinamos teoremos, analogiškos didžiųjų skaičių dėsniui (žr. pračio skyrelio p. 6), tačiau formuluojamos ir įrodomos, pakeičiant konvergavimą pagal tikimybę į konvergavimą su tikimybe 1.

Naudojant sustiprintą didžiųjų skaičių dėsnį, pračio skyrelio teoremą apie empirinių matų konvergavimą galima sustiprinti taip:

Glivenko-Kanteli teorema. Jei $X_1, \dots, X_n$ yra imtis iš populiacijos, kurios tikimybių skirstinys yra P , ir $n \rightarrow \infty$, tai

$$\sup_{B \in \mathcal{T}} |P_n^*(B) - P(B)| \xrightarrow{\text{su tikimybe 1}} 0,$$

kur $\mathcal{T}$ yra aibė pusintervalių $[a, b)$ su su baigtiniais arba begaliniais galais.

5. Šią teoremą empiriniams skirstiniams patogiau formuluoti kita forma.

Glivenko teorema. Jei $X_1, \dots, X_n$ yra imtis iš populiacijos, kurios skirstinio funkcija yra $F(x)$, o empirinis skirstinys yra $F_n^*(x)$, tai

$$P\left(\left\{\omega \mid \sup_{-\infty < x < \infty} |F_n^*(x) - F(x)| \rightarrow 0, \quad \text{kai } n \rightarrow \infty\right\}\right) = 1.$$

Iš čia darome išvadą, kad empirinis skirstinys $F_n^*(x)$ gerai aproksimuoja populiacijos skirstinį $F(x)$, kai n yra pakankamai didelis.

34. EMPIRINIŲ CHARAKTERISTIKŲ KONVERGAVIMAS Į JŲ TEORINIUS ANALOGUS

1. Jau pirmojoje vadovėlio dalyje susidūrėme su tokiais sąvokomis, kaip empirinis vidurkis, empirinė dispersija. Pats žodis *empirinis* reiškia *pagrįstas patirtimi*. Teorijoje *empirinėmis charakteristikomis* paprastai vadinami įvairūs išmatuojami funkcionalai nuo empirinio skirstinio.

Pavyzdžiui, *empiriniu k-tosios eilės momentu* vadinamas dydis

$$M_k^* = M_k^*(\mathbf{X}_n) = \int_{-\infty}^{\infty} x^k dF_n^*(x) = \frac{1}{n} \sum_{i=1}^n X_i^k,$$

kur, kaip paprastai, $\mathbf{X}_n = (X_1, \dots, X_n)$ yra nagrinėjama imtis.

Empiriniu centruotu k-tosios eilės momentu vadinamas dydis

$$S_k^* = S_k^*(\mathbf{X}_n) = \int_{-\infty}^{\infty} (x - M_1^*)^k dF_n^*(x) = \frac{1}{n} \sum_{i=1}^n (X_i - M_1^*)^k.$$

Empiriniam momentui M_1^* prigijo pažymėjimas $\bar{X}$ ir jis yra vadinamas *empiriniu vidurkiu*. Rusų kalba parašytuose vadovėliuose empirinis centruotas antros eilės momentas dažnai vadinamas *empirine dispersija*:

$$\bar{X} = M_1^* = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_2^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Šiame vadovėlyje *empirine* arba *imties dispersija* vadinama tokia statistika:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Apie tai buvo kalbama 7-ajame skyrelyje.

2. Daugelyje net palyginti paprastų kalkuliatorių yra numatyti klavišai empirinio vidurkio $\bar{X}$, empirinės dispersijos S^2 (arba empirinio standartinio nuokrypio S), empirinio antrojo momento M_2^* ir kt. apskaičiavimui. Siūlome skaitytojui savarankiškai susipažinti su tokio tipo kalkuliatorių naudojimo instrukcija ir praktinio pavyzdžio sprendimu įsitikinti, kad kalkuliatoriaus panaudojimas šiam tikslui yra gana efektyvus.

17 lentelė

Ūgis (cm)	Aštuntokų skaičius	Ūgis (cm)	Aštuntokų skaičius	Ūgis (cm)	Aštuntokų skaičius
143–145	1	158–160	120	173–175	64
146–148	2	161–163	181	176–178	28
149–151	8	164–166	201	179–181	10
152–154	26	167–169	170	182–184	3
155–157	65	170–172	120	185–187	1

Pavyzdys. Buvo atliktas 1000 berniukų ūgio matavimai. Atliktų matavimų rezultatai surašyti šioje lentelėje:

Apskaičiuoti vieno centimetro tikslumu empirinį vidurkį $\bar{X}$, empirinę dispersiją S^2 , empirinį centruotą momentą S_2^* ir empirinį antrosios eilės momentą M_2^* .

Atsakymas: $\bar{X} = 165,03 \approx 165$, $S^2 = 36,68 \approx 37$, $S_2^* = ((n - 1)/n)S^2 = 36,64 \approx 37$, $M_2^* = 27271,476 \approx 27271$.

3. Praeitame skyrelyje minėjome, kad sustiprinto didžiųjų skaičių dėsnio teoremos – tai teoremos, analogiškos didžiųjų skaičių dėsniai, tačiau pakeičiant konvergavimą pagal tikimybę į konvergavimą su tikimybe 1.

Jei yra žinoma, kad atsitiktinių dydžių sekoje kiekvienas iš šių dydžių turi vieną ir tą patį skirstinį, tai, pasirodo, sustiprintą didžiųjų skaičių dėsnį galima formuluoti, nereikalaujant dispersijos baigtinumo. Mes pateiksime šį dėsnį be įrodymo ir pasinaudosime juo įrodinėdami teoremą apie empirinių charakteristikų konvergavimą į jų teorinius analogus.

Sustiprintas didžiųjų skaičių dėsnis. Tegu $\{X_n\}$ – seka nepriklausomų vienodai pasiskirsčiusių atsitiktinių dydžių su baigtiniu vidurkiu $EX_n = a$. Tada

$$P\left(\left\{\omega \mid \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k(\omega) = a\right\}\right) = 1.$$

Pagaliau, funkcija $f : (-\infty, \infty) \rightarrow (-\infty, \infty)$ yra vadinama *Borelio funkcija*, jei ji yra mačioji funkcija σ -algebros atžvilgiu.

4. Suformuluosime ir įrodysime teoremą apie empirinių charakteristikų konvergavimą į jų teorinius analogus.

Teorema. Tegu $\mathbf{X}_n = (X_1, \dots, X_n)$ yra imtis iš populiacijos, kurios skirstinys yra F , o g yra Borelio funkcija tokia, kad

$$\mathbf{E}g(X_i) = \int_{-\infty}^{\infty} g(x) dF(x) < \infty.$$

Jei $S(\mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^n g(X_i)$, tai, kai $n \rightarrow \infty$,

$$S(\mathbf{X}_n) \xrightarrow{\text{su tikimybe 1}} \mathbf{E}g(X_i),$$

t.y. jei $G(F) = \int_{-\infty}^{\infty} g(x) dF(x)$, tai, kai $n \rightarrow \infty$,

$$G(F_n^*) \xrightarrow{\text{su tikimybe 1}} G(F).$$

Irodymas. Kadangi

$$S(\mathbf{X}_n) = \int_{-\infty}^{\infty} g(x) dF_n^*(x) = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

yra suma nepriklausomų vienodai pasiskirsčiusių atsitiktinių dydžių $g(X_i)$, kurių vidurkis yra

$$\mathbf{E}g(X_i) = \int_{-\infty}^{\infty} g(x) dF(x) < \infty,$$

tai pagal p. 3^o suformuluotą sustiprintą didžiųjų skaičių dėsnį

$$S(\mathbf{X}_n) \xrightarrow{\text{su tikimybe 1}} \mathbf{E}g(X_i).$$

O kadangi $S(\mathbf{X}_n) = G(F_n^*)$, $\mathbf{E}g(X_i) = G(F)$, tai šią formulę galima perrašyti taip:

$$G(F_n^*) \xrightarrow{\text{su tikimybe 1}} G(F).$$

Teorema įrodyta.

5. Iš šios teoremos gauname, kad absoliutūs ir centruoti empiriniai momentai su tikimybe 1 konverguoja į jų teorinius analogus:

$$M_k^* = \frac{1}{n} \sum_{j=1}^n X_j^k \xrightarrow{\text{su tikimybe 1}} \mathbf{E}X_1^k,$$

$$S_k^* = \frac{1}{n} \sum_{j=1}^n (X_j - M_1^*)^k \xrightarrow{\text{su tikimybe 1}} \mathbf{E}(X_1 - \mathbf{E}X_1)^k,$$

$$S^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 \xrightarrow{\text{su tikimybe 1}} \sigma^2 = \mathbf{E}(X_1 - \mathbf{E}X_1)^2.$$

6. Mes, tokiu būdu gavome svarbią išvadą:

Empirinis skirstinys ir plati jo funkcionalų klasė, augant imties dydžiui, konverguoja į atitinkamas teorines reikšmes.

35. PAGRINDINIS ĮVERTINIŲ KONSTRAVIMO METODAS. SUDERINTASIS ĮVERTINYS

1. Priminsime (žr. 12-ąjį skyrelį), kad pagal Lietuvos standartą *įvertinys* – tai statistika, kuria naudojamosi populiacijos parametrai įvertinti, o *įvertis* – tai įvertinio reikšmė, gauta iš konkrečios imties.

Nagrinėjant formaliai, terminas *įvertinys* išreiškia tą patį, ką ir terminas *statistika*. Tai bet kuri mačioji funkcija, kurios argumentas – imtis.

Tačiau nagrinėjant neformaliai, įvertinys – tai tokia statistika θ^* , kuri skirta naudoti vietoj nežinomo parametro θ .

Kitaip sakant, įvertinys θ^* yra tam tikra parametro θ aproksimacija imties pavidalu.

2. Įvertinio konstravimas – tai imties $\mathbf{X}_n = (X_1, X_2, \dots, X_n)$ funkcijos $\theta^* = \theta_n^*(\mathbf{X}_n)$ apibrėžimas visiems galimiems $n = 1, 2, \dots$

Beveik visų įvertinių pagrindas – tai teorinių skirstinių keitimo empiriniais skirstiniais metodas, pagrįstas teorema iš pračio skyrelio. Sutrumpintai šis metodas vadinamas *keitimo metodu*.

3. **Keitimo metodo aprašymas.** Tegų $\mathbf{X}_n = (X_1, \dots, X_n)$ yra imtis, paimta iš populiacijos su skirstiniu F .

Tegų, be to, nežinomas parametras θ išreiškiamas funkcionalu G nuo skirstinio F , t.y. $\theta = G(F)$.

Pavyzdžiui, $\theta = \int_{-\infty}^{\infty} x dF(x)$ arba $\theta = \int_{-\infty}^{\infty} (x-1)^2 dF(x)$.

Tegu, kaip ir anksčiau, F_n^* yra imties $\mathbf{X}_n = (X_1, \dots, X_n)$ empirinis skirstinys.

Tada *keitimo metodas* rekomenduoja pasirinkti įvertiniu θ^* funkciją

$$\theta^* = G(F_n^*),$$

t.y. funkcionalė $\theta = G(F)$ teorinį skirstinį F pakeisti empiriniu skirstiniu F_n^* . Tokie įvertiniai vadinami *keitimo įvertiniais*.

4. Šio metodo tikslingumas grindžiamas Glivenko teorema apie F_n^* konvergavimą į F , o taip pat praeito skyrelio teorema apie empirinių charakteristikų $G(F_n)$ konvergavimą į jų teorinius analogus $G(F)$.

Įvertiniai, konverguojantys pagal tikimybę į vieną ar kitą skirstinio parametą, vadinami *suderintaisiais įvertiniais*, o konverguojantys su tikimybe 1 – *stipriai suderintaisiais įvertiniais*. Užrašysime šiuos apibrėžimus detalesne forma.

5. Įvertis $\theta_n^* = \theta_n^*(\mathbf{X}_n)$ vadinamas parametro θ *suderintuoju įvertiniu*¹, jei

$$\theta_n^* = \theta_n^*(\mathbf{X}_n) \xrightarrow[\text{tikimybę}]{\text{pagal}} \theta.$$

Įvertis $\theta_n^* = \theta_n^*(\mathbf{X}_n)$ vadinamas *stipriai suderintuoju įvertiniu*, jei

$$\theta_n^* = \theta_n^*(\mathbf{X}_n) \xrightarrow[\text{su tikimybe 1}]{} \theta.$$

6. Skyrelio pabaigoje suformuluosime tokį rezultatą.

Teorema. Tegu

$$\theta = G(F) = \int_{-\infty}^{\infty} g(x) dF(x) < \infty,$$

kur F yra populiacijos skirstinys, o g – Borelio funkcija. Jei $\mathbf{X}_n = (X_1, \dots, X_n)$ yra šios populiacijos imtis, o F_n^* – šios imties empirinis skirstinys, tai $\theta^* = G(F_n^*)$ yra parametro θ stipriai suderintasis įvertinys.

Šios teoremos įrodymas yra elementarus, jei prisiminti praeito skyrelio teoremą. Faktiškai ji tėra pastarosios teoremos performulavimas.

¹ Angl. *consistent estimator*, rus. *состоятельная оценка*.

36. SILPNO KONVERGAVIMO SĄVOKA.

EMPIRINIŲ CHARAKTERISTIKŲ ASIMPTOTINIS NORMALUMAS

1. 14-ajame skyrelyje nagrinėjant empirinę centrinės ribinės teoremos interpretaciją matėme, kad jei imties dydis n yra pakankamai didelis, tai empirinis vidurkis $\bar{X}$ yra pakankamai gerai aproksimuojamas normaliuoju skirstiniu. Tiksliau, $(\bar{X} - \mu)/(\sigma\sqrt{n})$ skirstinys yra pakankamai gerai aproksimuojamas standartinio normaliojo atsitiktinio dydžio Z skirstiniu.

Jei toks rezultatas yra teisingas paprasčiausiam keitimo įvertiniui $\bar{X}$, tai natūralu panagrinėti ir bendresnį atvejį, vietoj $\bar{X}$ paimant bet koki keitimo įvertinį $G(F_n^*)$. Pasirodo, kad tam tikra prasme analogiškas atsakymas yra teisingas ir bendru atveju. Norint tuo įsitikinti, reikalingos kai kurios papildomos sąvokos.

2. Sakoma, kad atsitiktinių dydžių seka $\{X_n\}$ *silpnai konverguoja* į X , jei jų skirstinių seka $\{F_n\}$ konverguoja į X skirstinį $F(x)$ visuose jo tolydumo taškuose. Tai žymima

$$X_n \Rightarrow X \text{ arba } F_n \Rightarrow F.$$

Galima įrodyti, kad iš sekos konvergavimo pagal tikimybę (ir juo labiau iš konvergavimo su tikimybe 1) išplaukia šios sekos silpnas konvergavimas.

Remiantis bendrosiomis matematinės analizės koncepcijomis taip pat galima įrodyti, kad jei charakteristinių funkcijų seka $\{f_n\}$ konverguoja į charakteristinę funkciją f , tai atitinkama skirstinių seka $\{F_n\}$ konverguoja į skirstinį F jo tolydumo taškuose, t.y. $F_n \Rightarrow F$.

3. Pagaliau, mums prireiks centrinės ribinės teoremos vienodai pasiskirsčiusių dėmenų atvejui, kuri, naudojant charakteristinių funkcijų aparatą, nesunkiai įrodoma tikimybų teorijos kurse.

Centrinė ribinė teorema. Tegu $X_1, \dots, X_n$ – nepriklausomų vienodai pasiskirsčiusių atsitiktinių dydžių seka. Tegu, be to, $EX_1 = 0$ ir $DX_1 = 1$. Tada normuotų sumų $(X_1 + \dots + X_n)/\sqrt{n}$ charakteristinių funkcijų seka konverguoja į standartinio normaliojo dėsnio charakteristinę funkciją $\exp(-t^2/2)$.

Išvada. Jei išpildytos centrinės ribinės teoremos sąlygos, tai

$$(X_1 + \dots + X_n)/\sqrt{n} \Rightarrow Z.$$

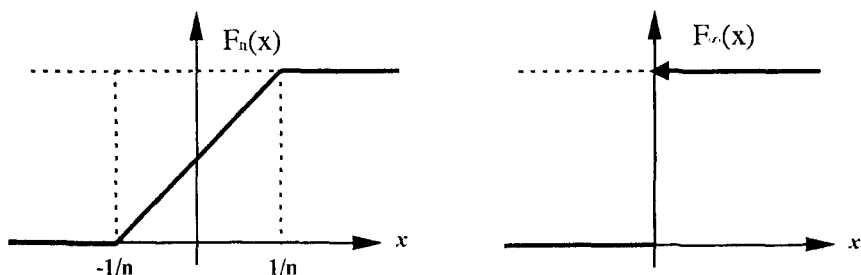
Išvados teisingumui įrodyti pakanka atsižvelgti į teiginį punkte 2 apie charakteristinių funkcijų sekos konvergavimo ir skirstinių sekos silpno konvergavimo sąryšį.

4. Kodėl pasirinktas *silpno* konvergavimo terminas? Gal todėl, kad vartojant vien terminą *konvergavimas*, paprastai turima galvoje konvergavimą *tolygiojoje metrikoje* ρ :

$$\rho(F, G) = \sup_x |F(x) - G(x)|.$$

O kadangi nesunku sukonstruoti skirstinių $\{F_n\}$ seką, silpnai konverguojančią į kažkokį skirstinį F_∞ , bet diverguojančią tolygiojoje metrikoje, tai pastaroji kažkuria prasme yra „stipresnė“, o pirmoji ne tokia „griežta“, „*silpnesnė*“.

Pateiksime tokios sekos pavyzdį. Iš tiesų, tegu $F_n(x)$ – tolygusis intervale $[-1/n, 1/n]$ skirstinys, o $F_\infty(x)$ – išsigimęs taške $x = 0$ skirstinys (45 pav.).



45 pav. Silpnai konverguojančios į išsigimusį skirstinį $F_\infty(x)$ tolygiųjų skirstinių sekos $\{F_n(x)\}$ pavyzdys, kuri diverguoja tolygiojoje metrikoje

Akivaizdu, kad visiems natūriniais n $\rho(F_n, F_\infty) \approx 1/2$. Taigi, $\rho(F_n, F_\infty) \not\rightarrow 0$, bet $F_n \Rightarrow F$.

5. Grįžkime prie uždavinio, suformuluoto 1 punkte, nagrinėjimo.

Parametro θ įvertinys $\theta^* = \theta^*(X_1, \dots, X_n)$ yra vadinamas *asimptotiškai normaliu* (su koeficientu $\sigma^2 \geq 0$), jei statistika $(\theta^* - \theta)\sqrt{n}$ silpnai konverguoja į normalųjį atsitiktinį dydį su vidurkiu 0 ir dispersija σ^2 . Galima ir tokia terminologija: įvertinys θ^* yra *asimptotiškai normalus* (su parametrais $(\theta, \sigma/n)$).

Teorema. Tegu $\mathbf{X}_n = (X_1, \dots, X_n)$ yra populiacijos su skirstiniu $F(x)$ imtis ir tegu

$$S_n = S_n(\mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^n g(X_i),$$

kur $g(x)$ – bet kokia Borelio funkcija tokia, kad

$$\int_{-\infty}^{\infty} g^2(x) dF(x) < \infty.$$

Tada įvertinys $S_n(\mathbf{X}_n)$ yra asimptotiškai normalus su parametrais $(\mu, \sigma^2/n)$, kur

$$\mu = \int_{-\infty}^{\infty} g(x) dF(x), \quad \sigma^2 = \int_{-\infty}^{\infty} (g(x) - \mu)^2 dF(x).$$

6. Įrodymas. Atsitiktinių dažnių sekai

$$\left\{ \frac{g(X_1) - \mu}{\sigma}, \frac{g(X_2) - \mu}{\sigma}, \dots, \frac{g(X_n) - \mu}{\sigma}, \dots \right\}$$

galime taikyti centrinę ribinę teoremą iš 3 punkto, nes

$$\begin{aligned} \mathbf{E} \frac{g(X_n) - \mu}{\sigma} &= \frac{1}{\sigma} \mathbf{E} g(X_n) - \frac{\mu}{\sigma} = 0, \\ \mathbf{D} \frac{g(X_n) - \mu}{\sigma} &= \frac{1}{\sigma^2} \mathbf{D} g(X_n) = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

Pagal šios teoremos išvadą,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{g(X_j) - \mu}{\sigma} \Rightarrow Z,$$

t.y.

$$\sqrt{n} \frac{1}{\sigma n} \sum_{j=1}^n g(X_j) - \frac{\mu}{\sigma/\sqrt{n}} \Rightarrow Z.$$

Iš čia gauname, kad

$$\frac{1}{\sigma/\sqrt{n}} (S_n - \mu) \Rightarrow Z,$$

t.y. įvertinys S_n yra *asimptotiškai normalus* su parametrais $(\mu, \sigma^2/n)$.

Teorema įrodyta.

37. MOMENTŲ METODAS KAIP KEITIMO METODO REALIZACIJA

1. Tegų $\{P_\theta | \theta \in \Theta\}$ yra mums žinoma tikimybinių skirstinių, priklausančių nuo nežinomo parametro θ , aibė. Parametras θ gali būti ir skaliarinis, ir vektorinis. Pavyzdžiui, jei kalbama apie normaliųjų skirstinių aibę $\{\Phi_\theta | \theta \in \Theta\}$, tai $\theta = (\alpha, \sigma^2)$ yra dvimatis vektorius, o Θ yra arba pusplokštumė $\{(\alpha, \sigma) | -\infty < \alpha < \infty, \sigma \geq 0\}$, arba, priklausomai nuo uždavinio sąlygų, jos dalis.

2. Bet kokios statistikos $S = S(X)$ matematinę viltį ir dispersiją pagal tikimybinį skirstinį P_θ žymėsime $E_\theta S$, $D_\theta S$.

Tegų $g(x)$ yra Borelio funkcija tokia, kad parametro θ funkcija $m(\theta)$,

$$m(\theta) = E_\theta g(X) = \int_{-\infty}^{\infty} g(x) P_\theta(dx). \quad (1)$$

būtų monotoningė ir tolydi. Čia X yra atsitiktinis dydis, turintis tikimybinį skirstinį P_θ , o θ yra vienamatis parametras.

Tarkime, kad funkcijos $m(\theta)$ reikšmių sritis $m(\Theta)$ turi tą pačią prigimtį, kaip ir Θ . Jei, pavyzdžiui, Θ yra realios ašies atkarpa, tai $m(\Theta)$ taip pat yra atkarpa.

Kadangi $m(\theta)$ yra monotoningė ir tolydi, tai lygtis $m(\theta) = t$ srityje $m(\Theta)$ turi vienareikšmį sprendinį $\theta = m^{-1}(t)$.

Tai reiškia, kad sąryšį (1) galima perrašyti taip:

$$\theta = m^{-1} \left(\int_{-\infty}^{\infty} g(x) P_\theta(dx) \right). \quad (2)$$

4. Pastebėsime, kad *keitimo metodas* rekomenduoja parametru $\theta = G(F)$ įvertinui pasirinkti funkciją $\theta^* = G(F_n^*)$. Todėl formulės (2) pagrindu pasirenkame

$$\theta^* = m^{-1} \left(\int_{-\infty}^{\infty} g(x) dF_n^*(x) \right) = m^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(X_i) \right), \quad (3)$$

kur, kaip paprastai, $X_1, \dots, X_n$ yra imtis iš populiacijos su tikimybinio skirstiniu P_θ .

Įvertinys θ^* , išreiškiamas (3) formule, yra vadinamas *įvertiniu, gautu momentų metodu*.

Tokią pavadinimą galima paaiškinti tuo, kad funkcija $g(x)$ paprastai parenkama taip:

$$g(x) = x \quad \text{arba} \quad g(x) = x^k \quad (k > 1).$$

Tokiais atvejais įvertinys (3) yra lygties

$$m(\theta^*) = \frac{1}{n} \sum_{i=1}^n g(X_i) \quad (4)$$

sprendinys, o pastaroji yra paprastųjų momentų lygties (1) empirinis analogas.

5. Pavyzdys. Tegu $X_1, \dots, X_n$ yra imtis iš populiacijos, turinčios *eksponentinį skirstinį* $E(x)$, t.y. $E(x) = 1 - \exp(-\lambda x)$, kai $x \geq 0$ ir $E(x) \equiv 0$, kai $x < 0$. Čia $\lambda > 0$ yra parametras. Tada

$$\begin{aligned} EX_1^k &= \int_{-\infty}^{\infty} x^k dE(x) = \int_0^{\infty} x^k \lambda \exp(-\lambda x) dx = \\ &= \frac{1}{\lambda^k} \int_0^{\infty} t^k \exp(-t) dt = k!/\lambda^k. \end{aligned}$$

Norint sukonstruoti parametro λ įvertinį λ^* momentų metodu, pakanka paimti $g(x) = x$ ir pastebėti, kad tokiu atveju formulėje (1)

$$m(\theta) = E_\theta X = 1!/\theta^1 = 1/\theta.$$

Taigi, iš formulės (4) turime, kad

$$1/\lambda^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}, \quad \text{t.y.} \quad \lambda^* = 1/\bar{X}.$$

38. PAPRASČIAUSIAS STATISTINIŲ ĮVERTINIŲ PALYGINIMAS. SISTEMINGOJI PAKLAIDA

1. Matėme, kad net vienam parametrai θ egzistuoja be galo daug statistinių įvertinių $\theta_n^*(X_1, \dots, X_n)$. Kuris iš šių įvertinių geriausiai reprezentuoja nežinomąjį parametą θ , geriausiai jį aproksimuoja, yra arčiausiai šio parametro? Suprantama, pirmiausia iškyla klausimas, ką reiškia sąvokos *geriausiai reprezentuoja, geriausiai aproksimuoja, arčiausiai*.

O gal, nagrinėjant parametro θ įvertinius θ_n^* , normalu tikėtis, kad skirtumo

$$\theta_n^*(X_1, \dots, X_n) - \theta \quad (1)$$

skirstinys būtų pakankamai sukoncentruotas apie nulį?

2. Bet praktikoje dažnai nagrinėjamas ne atstiktinio dydžio (1) skirstinys, o pirmieji du momentai. Tokiais atvejais skirtumo (1) kokybė charakterizuojama poslinkiu

$$\mathbf{E}_\theta \theta_n^*(X_1, \dots, X_n) - \theta \quad (2)$$

ir dispersija

$$\mathbf{E}_\theta (\theta_n^*(X_1, \dots, X_n) - \mathbf{E}_\theta \theta_n^*)^2. \quad (3)$$

Įvertinys $\theta_n^*(X_1, \dots, X_n)$ vadinamas parametro θ nepaslinktuoju įvertiniu¹, jei skirtumas (2) lygus nuliui.

Tokiu atveju dažnai yra sakoma, kad nedaroma sistemingoji paklaida.

3. Kadangi kvadratinio nuokrypio vidurkis $\mathbf{E}_\theta (\theta_n^* - \theta)^2$, kaip matėme pirmojoje vadovėlio dalyje, yra pakankamai geras įvertinio θ_n sklaidos matas, o nepaslinktojo įvertinio atveju $\theta = \mathbf{E}_\theta \theta_n^*$, tai yra teisinga tokia išvada:

Nepaslinktųjų įvertinių kokybė nusakoma jų dispersijos (3) reikšme – kuo ši reikšmė mažesnė, tuo įvertinys yra geresnis.

4. Ne visi mūsų nagrinėti konkretūs statistiniai įvertiniai yra nepaslinktieji įvertiniai. Pavyzdžiui, centruotas antrasis empirinis momentas

$$S_2^* = \frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

yra nors ir suderintasis, bet nėra nepaslinktasis atsitiktinės imties su dispersija σ^2 įvertinys.

¹ Angl. *unbiased estimator*, rus. *несмещенная оценка*.

Iš tiesų, kadangi $E(X_k - \bar{X}) = 0$ ir

$$\begin{aligned} E(X_k - \bar{X})^2 &= D(X_k - \bar{X}) = D\left\{\left(1 - \frac{1}{n}\right)X_k - \frac{1}{n}\sum_{r \neq k} X_r\right\} = \\ &= \left(1 - \frac{1}{n}\right)^2 D X_k + \frac{1}{n^2} \sum_{r \neq k} D X_r = \frac{n-1}{n} \sigma^2, \end{aligned}$$

tai

$$ES_2^* = \frac{1}{n} \sum_{k=1}^n E(X_k - \bar{X})^2 = \frac{n-1}{n} \sigma^2 = \left(1 - \frac{1}{n}\right) \sigma^2.$$

Tai reiškia, kad pasirinkus $\theta_n^* = S_2^*$, $\theta = \sigma^2$, skirtumas (2) nėra lygus nuliui, t.y. S_2^* nėra dispersijos σ^2 nepaslinktasis įvertinys. Štai kodėl *empirine dispersija* mes pavadino ne statistiką S_2^* , o šiek tiek „pataisytą“ jos variantą $\frac{n}{n-1} S_2^*$, t.y. *empirine dispersija* pavadino nepaslinktąjį įvertinį S^2 ,

$$S^2 = \frac{n}{n-1} S_2^* = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2.$$

5. Skyrelio pabaigoje be komentarų pateiksime kelias sąvokas iš Lietuvos standarto:

įvertinio paklaida – tai *parametro vertinimo paklaida* $(\theta^* - \theta)$; čia θ^* – įvertinys, o θ – vertinamasis parametras;

įvertinio poslinkis – tai *įvertinio vidurkio ir vertinamojo parametro skirtumas*;

nepaslinktasis įvertinys – tai *įvertinys, kurio poslinkis lygus nuliui*;

standartinė paklaida – tai *įvertinio standartinis nuokrypis*.

39. DOMINUOJANTYSIS MATAS. MINIMIZACIJOS LEMA

1. Nagrinėsime parametrinę matų šeimą $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$. Tarkime, kad ši šeima tenkina tokias dvi savybes:

(A1) $P_{\theta_1} \neq P_{\theta_2}$, kai $\theta_1 \neq \theta_2$;

(A2) egzistuoja σ -baigtinis matas ν toks, kad visi $P_\theta \in \mathcal{P}$ turi šio mato atžvilgiu tankį

$$f_\theta(x) = \frac{dP_\theta}{d\nu}(x), \quad \text{t. y. } P_\theta(B) = \int_B f_\theta(x) \nu(dx)$$

bet kuriai Borelio aibei B .

Jei parametrinis matų šeimos $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ elementai tenkina sąlygas (A1) ir (A2), tai sakoma, kad matas ν dominuoja aibėje $\mathcal{P}$.

2. Absoliučiai tolygiųjų skirstinių aibėje, t.y. skirstinių, turinčių tankius, dominuojantysis matas ν yra paprastas Lebego matas. Pavyzdžiui, standartinio normaliojo skirstinio $\Phi(u)$ atveju turime, kad

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u \exp(-x^2/2) dx, \quad \text{t.y. } \nu(dx) = dx.$$

3. Diskrečiųjų skirstinių aibėje ν yra taip vadinamas *skaičiuojantysis matas* (detaliau žr. [1], sk. 2, §2 ir §6), o tankis $f_\theta(x)$ sutampa su tikimybės masės funkcija. Pastarąsias nagrinėjome 5-ajame skyrelyje.

Trumpai pakomentuosime diskretųjį atvejį. Kas yra tankis F_θ tolygiuoju atveju? Tai skirstinio F_θ išvestinė, ir todėl

$$f_\theta(x)dx = dF_\theta(x) = \lim_{\Delta x \downarrow 0+} (F_\theta(x + \Delta x) - F_\theta(x)).$$

Tankio analogas diskrečiuoju atveju – tai trūkio didumas p taške x : $F(x + \Delta x) - F(x) = p = \mathbf{P}_\theta(X = x)$.

Vienas iš dažniausiai sutinkamų diskrečiųjų skirstinių – tai *Puasono skirstinys*

$$\Pi_\lambda(x) = \sum_{k=0}^{[x]} \frac{\lambda^k}{k!} \exp(-\lambda), \quad \text{kai } x \geq 0,$$

ir $\Pi_\lambda(x) = 0$, kai $x < 0$. Čia λ – teigiamas parametras, o $[x]$ – reikšmės x sveikoji dalis. Skaičiuojantysis matas $\nu(x)$ šiuo atveju pilnai apibrėžiamas tokia formule:

$$\nu(x) = \begin{cases} [x] + 1, & \text{jei } x \geq 0, \\ 0, & \text{jei } x < 0. \end{cases}$$

4. Jei matas ν dominuoja aibėje $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, tai tarp bet kurių dviejų aibės $\mathcal{P}$ elementų F ir G galima apibrėžti *atstumą*, pavyzdžiui, tokiu būdu:

$$d(F, G) = - \int_{-\infty}^{\infty} g(x) \ln f(x) \nu(dx); \quad (1)$$

čia

$$F(B) = \int_B f(x) \nu(dx), \quad G(B) = \int_B g(x) \nu(dx).$$

Mes nekonkretizuojame sąvokos *atstumas*, nors intuityviai norėtusi, kad funkcija su tokiu pavadinimu būtų neneigiama. O kadangi tankis – neneigiama funkcija, kurios integralas nuo $-\infty$ iki $+\infty$ yra lygus 1, tai formulėje (1) prieš integralo ženklą stovi minuso ženklas. Punktuose 2 ir 3 išnagrinėtais atvejais tankis neviršija 1, todėl logaritmas nuo šio tankio yra neigiamas, taigi formulės (1) dešinė pusė yra teigiama.

Beje, *atstumas*, apibrėžiamas formule (1), neturi kai kurių atstumo, suprantamo buitine prasme, savybių. Pavyzdžiui, *atstumas* (1) nėra simetrinis, $d(F, F) \neq 0$ ir panašiai.

5. Sekančiame skyrelyje mums reikalinga nelygybė

$$d(F, F) \leq d(F, G). \quad (2)$$

Ši nelygybė parodo, kad atstumas $d(F, G)$ yra minimalus tada, kai $F = G$. Nelygybė (2) yra įrodoma sekančioje lemoje.

Minimizacijos lema. Tegu f ir g yra du tikimybiniai tankiai mato atžvilgiu. Tada

$$\int_{-\infty}^{\infty} f(x) \ln f(x) \nu(dx) \geq \int_{-\infty}^{\infty} f(x) \ln g(x) \nu(dx),$$

jei abu šie integralai yra baigtiniai.

Įrodymas. Mums reikia įrodyti, kad

$$\int_{-\infty}^{\infty} f(x) \ln \frac{g(x)}{f(x)} \nu(dx) \leq 0.$$

Kadangi $\ln(1+x) \leq x \quad \forall x \geq -1$ ir lygybė galima tik kai $x = 0$, tai

$$\ln \frac{g(x)}{f(x)} = \ln \left(1 + \left(\frac{g(x)}{f(x)} - 1 \right) \right) \leq \frac{g(x)}{f(x)} - 1.$$

Todėl

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) \ln \frac{g(x)}{f(x)} \nu(dx) &\geq \int_{-\infty}^{\infty} f(x) \left(\frac{g(x)}{f(x)} - 1 \right) \nu(dx) = \\ &= \int_{-\infty}^{\infty} g(x) \nu(dx) - \int_{-\infty}^{\infty} f(x) \nu(dx) = 1 - 1 = 0. \end{aligned}$$

Lema įrodyta.

Išvada. Jei matas ν dominuoja aibėje $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, tai bet kuriems šios aibės elementams P_θ ir P_{θ_1}

$$d(P_\theta, P_\theta) \leq d(P_\theta, P_{\theta_1}),$$

jei tik šie atstumai yra baigtiniai.

Įrodymas akivaizdus, nes faktiškai ši išvada yra lemos reformulavimas.

40. DIDŽIAUSIO TIKĖTINUMO ĮVERTINYS IR JO APSKAIČIAVIMAS KONKREČIAIS ATVEJAIS

1. Atstumą d apibrėšime kaip praeitame skyrelyje. O būtent, jei matas ν dominuoja aibėje $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, tai tarp bet kurių dviejų aibės $\mathcal{P}$ elementų P_{θ_i} ir P_{θ_j} atstumą d apibrėžiame tokiu būdu:

$$d(P_{\theta_i}, P_{\theta_j}) = - \int_{-\infty}^{\infty} f_{\theta_j}(x) \ln f_{\theta_i}(x) \nu(dx);$$

čia $P_{\theta_i}(B) = \int_B f_{\theta_i}(x) \nu(dx)$, $P_{\theta_j}(B) = \int_B f_{\theta_j}(x) \nu(dx)$ bet kuriai Borelio aibei B .

2. Pagal lemos išvadą, bet kuriems $\theta \in \Theta$ ir $\theta_0 \in \Theta$ teisinga nelygybė

$$d(P_\theta, P_{\theta_0}) \geq d(P_{\theta_0}, P_{\theta_0}).$$

Tai reiškia, kad

$$\min_{\theta \in \Theta} d(P_\theta, P_{\theta_0}) \geq d(P_{\theta_0}, P_{\theta_0}).$$

Šis minimumas egzistuoja, nes, pagal prielaidą, $\theta_0 \in \Theta$. Kadangi

$$\min_{\theta_0 \in \Theta} \min_{\theta \in \Theta} d(P_\theta, P_{\theta_0}) \geq \min_{\theta_0 \in \Theta} d(P_{\theta_0}, P_{\theta_0}),$$

tai, norint rasti minimumą pagal atstumą d visoje klasėje $\mathcal{P}$, mes turime rasti

$$\min_{\theta_0 \in \Theta} d(P_{\theta_0}, P_{\theta_0}). \quad (1)$$

Šis minimumo taškas turi ypatingą reikšmę statistikoje, o pats atstumas d betarpiškai susijęs su taip vadinamu Kulback-Leibler atstumu tarp skirstinių. Mes neanalizuosime, kodėl šis taškas ypatingas. Toliau mus domins tik tai, kaip rasti šį tašką.

Taigi, ką reiškia rasti minimumą (1)? Tai reiškia rasti maksimumą

$$\max_{\theta \in \Theta} \int \ln f_\theta(x) dP_\theta(x). \quad (2)$$

3. Šio maksimumo įvertiniu keitimo metodas rekomenduoja pasirinkti statistiką, gaunamą formulėje (2) pakeitus teorinį skirstinį P_θ empiriniu skirstiniu P_n^* :

$$\max_{\theta} \int_{-\infty}^{\infty} \ln f_\theta P_n^*(dx) = \max_{\theta} \frac{1}{n} \sum_{i=1}^n \ln f_\theta(X_i). \quad (3)$$

Apibrėžimas. Jei egzistuoja statistika $\theta^* = \theta^*(X_1, \dots, X_n)$, tenkinanti sąlygą

$$L(\theta^*) = \max_{\theta} \sum_{i=1}^n \ln f_\theta(X_i), \quad (4)$$

tai ji yra vadinama didžiausio tikėtinumo įvertiniu¹.

4. Iš kur toks pavadinimas – didžiausio tikėtinumo įvertinys?

Jei nagrinėjamas diskretusis atvejis, tai tankis $f_\theta(x)$ sutampa su tikimybe $P_\theta(X_j = x)$. Kadangi bet kurios teigiamos funkcijos ir jos logaritmo ekstremumai pasiekiami vienoduose taškuose, tai įvertis $\theta = \theta(x_1, \dots, x_n)$, apskaičiuotas pagal formulę (4), tuo pačiu maksimizuoja funkciją

$$f_\theta(x_1)f_\theta(x_2)\dots f_\theta(x_n) = \prod_{j=1}^n P_\theta(X_j = x_j).$$

¹ Angl. *maximum likelihood estimator*; rus. *оценка максимального правдоподобия*.

O ši tikimybė ir yra maksimali įvykio $\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$ pasirodymo tikimybė.

5. Priminsime, kad mes žymėjome $\mathbf{X}_n = (X_1, \dots, X_n)$. Funkcija

$$f_{\theta}(\mathbf{X}_n) = \prod_{i=1}^n f_{\theta}(X_i)$$

kaip θ funkcija vadinama *tikėtinumo funkcija*, o funkcija

$$L(\theta) = L(\mathbf{X}_n, \theta) = \ln f_{\theta}(\mathbf{X}_n) = \sum_{i=1}^n l(X_i, \theta)$$

(kur $l(x, \theta) = \ln f_{\theta}(x)$) vadinama *logaritmine tikėtinumo funkcija*.

Jei Θ yra n -matis intervalas, funkcija $l(\theta) = l(\mathbf{X}_n, \theta)$ turi maksimumą jo viduje ir yra diferencijuojama $\theta_1, \dots, \theta_n$ atžvilgiu, tai jos maksimumo taške

$$\frac{\partial L(\theta)}{\partial \theta_j} = 0, \quad j = 1, \dots, n.$$

Vadinasi, didžiausio tikėtinumo įvertinys yra šių, taip vadinamųjų *didžiausio tikėtinumo, lygčių* sprendinys. Dažnai tos lygtys turi tik vieną sprendinį.

6. Apskaičiuosime didžiausio tikėtinumo įvertinius, kai nagrinėjama populiacija turi normalųjį ar gama skirstinius.

Normaliojo dėsnio (su parametrais μ ir σ) *tankis* yra

$$\varphi_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -(x - \mu)^2 / (2\sigma^2) \right\}, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

Pažymėję $\theta = (\mu, \sigma)$, šiuo atveju gauname, kad

$$f_{\theta}(\mathbf{X}_n) = (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 / (2\sigma^2) \right\},$$

$$L(\theta) = L(\mathbf{X}_n, \theta) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Kadangi logaritmas yra monotoninė funkcija, tai, kaip buvo minėta, f_{θ} ir $L(\theta)$ maksimumą pasiekia tame pačiame taške θ . Šiame taške turi būti tenkinamos dvi *didžiausio tikėtinumo lygtys*:

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) = 0,$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (X_i - \mu)^2 = 0.$$

Iš čia nesunkiai gauname, kad

$$\hat{\mu}^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

$$(\hat{\sigma}^2)^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_2^*.$$

Čia ir toliau simboliu $\hat{\cdot}$ žymimas didžiausio tikėtinoumo įvertis. Priminsime, kad 34-ajame skyrelyje S_2^* žymėjome empirinį centruotą antros eilės momentą. Beje, nesunku patikrinti, kad šiame dvimačiame taške funkcija L tikrai pasiekia maksimumą.

7. Išnagrinėsime *gama skirstinį*, kurio *tankis* apibrėžiamas taip:

$$\gamma_\alpha(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, \quad x \geq 0, \quad \alpha > 0.$$

Tarkime, kad parametras λ yra žinomas. Tada

$$f_\alpha(\mathbf{X}_n) = \frac{\alpha^{n\lambda}}{(\Gamma(\lambda))^n} \prod_{i=1}^n X_i^{\lambda-1} \exp\left(-\alpha \sum_{i=1}^n X_i\right),$$

$$L(\alpha) = L(\mathbf{X}_n, \alpha) = n\lambda \ln \alpha - n \ln \Gamma(\lambda) + (\lambda - 1) \sum_{i=1}^n \ln X_i - \alpha \sum_{i=1}^n X_i,$$

$$\frac{\partial L}{\partial \alpha} = \frac{n\lambda}{\alpha} - \sum_{i=1}^n X_i = 0, \quad \hat{\alpha}^* = \lambda / \bar{X}.$$

41. PAPERASTOJI IR SUDĖTINĖ HIPOTEZĖS. HIPOTEZIŲ TIPAI. PRAKTINIŲ IR TEORINIŲ REZULTATŲ SUDERINAMUMAS

1. Hipotezių formulavimas ir hipotezių tikrinimo procedūros nagrinėjimas buvo trumpai apžvelgtas 19, 20 ir 21 skyreliuose. Pratešime šį nagrinėjimą toliau.

Minėjome, jog gana dažnai yra žinoma, kad stebimo atsitiktinio dydžio skirstinys priklauso šeimai $\mathcal{P} = \{F_\theta | \theta \in \Theta\}$, tačiau konkreti parametro θ reikšmė yra nežinoma. Pavyzdžiui, $\mathcal{P}$ gali būti normaliųjų skirstinių su vidurkiu 0 ir su nežinoma dispersija σ^2 šeima. Hipotezės apie parametro reikšmės yra

vadinamos parametrinėmis hipotezėmis, o jų pavyzdžius mes nagrinėjome pirmojoje vadovėlio dalyje.

2. Jei hipotezė H_0 vienareikšmiškai fiksuoja parametro θ reikšmę θ_0 , t.y. hipotezė H_0 vienareikšmiškai nusako *hipotezinį skirstinį*¹, tai ji yra vadinama *paprastąja hipoteze*². Priešingu atveju hipotezė vadinama *sudėtine hipoteze*³.

O štai Lietuvos standarte pateikiami apibrėžimai ir pavyzdžiai. **Paprastoji hipotezė** – tai hipotezė, kuri visiškai nusako populiacijos skirstinį. Hipotezė, kuri nevysiškai nusako populiacijos skirstinį, vadinama **sudėtine hipoteze**. Dažniausiai sudėtinę hipotezę sudaro be galo daug paprastųjų hipotezių.

Pavyzdžiui, normaliojo skirstinio atveju hipotezė $\mu = \mu_0$ yra paprastoji, jeigu populiacijos standartinis nuokrypis yra žinomas, bet sudėtinė, jeigu jis yra nežinomas.

3. Dabar kalbėsime apie hipotezių tikrinimo uždavinius, kai šeimos $\mathcal{P}$ skirstinių funkcinė išraiška nėra žinoma.

Tiksliau, tarkime, kad imtis paimta iš populiacijos skirstinio F ir yra žinoma, kad F priklauso tam tikrai skirstinių šeimai $\mathcal{T}$. Ši šeima gali būti, pavyzdžiui, tolydžių, diskrečių ar pan. skirstinių aibė.

4. *Neparametrinė hipoteze* H vadiname teiginį tipo „skirstinys F priklauso aibei $\mathcal{T}$ “

Suformuluosime keletą būdingesnių neparametrinės statistikos *hipotezių tipų*.

Suderinamumo hipotezės. Tarkime, kad $X_1, \dots, X_n$ yra imtis iš populiacijos, kurios skirstinys yra $F(x)$. *Suderinamumo hipoteze* vadiname paprastąją hipotezę $H : F(x) \equiv F_0(x)$. Čia $F_0(x)$ yra visiškai nusakytas vienintelis aibės $\mathcal{T}$ elementas, t.y., aibė $\mathcal{T}$ susideda iš vienintelio skirstinio $F_0(x)$. Šis apibrėžimas dažnai yra praplečiamas tokiu būdu. *Suderinamumo hipoteze* taip pat vadiname sudėtinę hipotezę H ,

$$H : F(x) \in \mathcal{T}, \quad \text{kur} \quad \mathcal{T}_0 = \{F_\theta : \theta \in \Xi\}.$$

Pavyzdžiui, $\mathcal{T}_0$ yra normaliųjų, Puasono ar pan. skirstinių aibė.

Nepriklausomumo hipotezės. Tarkime, kad (X_{1i}, X_{2i}) , $i = 1, 2, \dots, n$ yra imtis iš dvimačio atsitiktinio vektoriaus (X_1, X_2) populiacijos.

¹ Angl. *hypothesized distribution*, rus. *зупотетическое распределение*.

² Angl. *simple hypothesis*, rus. *простая зупотеза*.

³ Angl. *composite hypothesis*, rus. *сложная зупотеза*.

Reikia patikrinti hipotezę $H : F(x_1, x_2) \in T$, kur

(A) $F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2)$, t.y. F yra vektoriaus (X_1, X_2) skirstinys;

(B) $T = \{\tilde{F}(x_1, x_2) \mid \tilde{F}(x_1, x_2) = \tilde{F}(x_1, \infty) \cdot \tilde{F}(\infty, x_2), \forall (x_1, x_2) \in \mathbb{R}^2\}$, t.y. T yra vektorių su nepriklausomomis komponentėmis skirstinių aibė.

Vienodo pasiskirstymo hipotezės. Kaip ir anksčiau, tarkime, kad (X_{1i}, X_{2i}) , $i = 1, 2, \dots, n$ yra imtis iš dvimačio atsitiktinio vektoriaus (X_1, X_2) populiacijos.

Reikia patikrinti hipotezę $H_0 : F(x_1, x_2) \in T$, kur $T = \{P(X_1 < x_1, X_2 < x_2) \mid P(X_1 < x) = P(X_2 < x)\}, \forall x \in \mathbb{R}^1$, t.y. T yra vektorių, kurių komponentių skirstiniai sutampa, skirstinių aibė.

5. Hipotezės sąvoka, jų tipai ir pan., žinoma, yra matematinės teorijos dalis. Tačiau matematinė teorija yra abstraktaus mąstymo rezultatas. Tam, kad teorinis modelis būtų praktiškai naudingas, yra būtina, kad jis būtų suderintas su jo empiriniu analogu.

Pavyzdžiui, jei mūsų teoriniame modelyje įvykio tikimybė yra maža, tai šis modelis yra praktiškai naudingas tik tuo atveju, jei eksperimente, kuris yra nagrinėjamo teorinio modelio praktinis analogas, šis įvykis yra realizuojamas tik atitinkamai mažai kartų pakankamai ilgoje šio eksperimento kartojimų sekoje.

Tai reiškia, jog mes turime būti praktiškai tikri, kad, atliekant eksperimentą vieną kartą, nagrinėjamas įvykis neįvyks.

Analogiškai, jei įvykio tikimybė artima vienetui, tai mes turime būti praktiškai tikri, kad šis įvykis įvyks, atlikus vienkartinę eksperimento realizaciją.

6. Daugeliu atvejų praktinių duomenų ir teorinių rezultatų *suderinamumo uždavinys* formuluojamas tokiu būdu. Turime stebimo atsitiktinio dydžio imtį, kurios dydis yra n , ir norime sužinoti, ar šio atsitiktinio dydžio skirstinys turi tam tikras savybes.

Pavyzdžiui, mus gali dominti klausimas, ar šis skirstinys yra normalusis su parametrais $m = 0$ ir $\sigma = 1$ (parametrinė hipotezė). Kitu atveju mums gali būti žinoma kažkokių skirstinių klasė (pavyzdžiui, vektorių su nepriklausomomis komponentėmis skirstinių klasė) ir gali dominti klausimas, ar mūsų imtis gauta iš populiacijos, kurios skirstinys priklauso šiai klasei (neparametrinė hipotezė).

7. Išnagrinėsime konkretų – pradžiai labai supaprastintą – praktinių duomenų ir teorinių rezultatų suderinamumo uždavinį.

Tarkime, kad norime patikrinti tokią nulinę hipotezę –

H_1 : Lietuvoje konservatorius remia 13% gyventojų, LDDP – 12% gyventojų, krikščionių demokratų partiją – 10%, kitas partijas arba dar neapsisprendę – 65% gyventojų.

Tarkime, be to, kad norėdami patikrinti šią hipotezę, mes apklausėme 1000 atsitiktinai parinktų Lietuvos gyventojų. Suformuluotos hipotezės pagrindu mes tikimės, kad apklausos rezultatai bus tokie: konservatorius parems 130, darbiečius – 120, krikščionis demokratų – 100, kitus – 650 apklaustųjų.

Jei, atlikę apklausą, mes pastebime, kad konservatorius parėmė 115, darbiečius – 115, krikščionis demokratų – 101, kitus – 669 apklaustieji, mes pasiruošę priimti suformuluotą hipotezę H_1 , nes skirtumas tarp duomenų, kuriuos mes tikėjomės gauti, ir duomenų, kuriuos mes realiai gavome, yra nedidelis. Tai reiškia, kad šiuo atveju praktiniai duomenys ir teoriniai rezultatai yra suderinami.

Tačiau jei, atlikę apklausą, mes gavome, kad konservatorius parėmė 30, darbiečius – 100, krikščionis demokratų – 100, o kitus – 770 apklaustųjų, tai šie duomenys prieštarauja hipotezės H_1 priėmimui. Tai, savo ruožtu, reiškia, kad tokiu atveju praktiniai duomenys ir teoriniai rezultatai yra nesuderinami.

Statistinė procedūra, leidžianti nustatyti, ar praktiniai duomenys ir teorinė hipotezė yra suderinami, vadinama *statistiniu kriterijumi* arba *testu*. Plačiau apie tai sekančiame skyrelyje.

42. SKIRSTINIO ATITIKIMAS. STATISTINIS KRITERIJUS (TESTAS). KRITINĖ SRITIS

1. Praeitame skyrelyje nagrinėjome praktinių duomenų ir teorinių rezultatų atitikimo (suderinamumo) klausimus. Pagal Lietuvos standartą, *skirstinio atitikimas*¹ – tai stebimo *skirstinio* ir *aprioriškai pasirinkto teorinio skirstinio atitikimas*.

Pereiname prie konkrečių skirstinio atitikimo problemų nagrinėjimo.

2. Patikrinsime hipotezę apie tai, kad mūsų imtis yra gauta iš populiacijos, kurios skirstinys yra $F(x)$ – žinoma mums funkcija (pavyzdžiui, konkretus eksponentinis skirstinys).

Tarkime, kad mūsų hipotezė yra teisinga. Tokiu atveju pagal Glivenko teoremą (žr. 33 skyrelį) empirinis skirstinys F_n^* gerai aproksimuoja populiacijos

¹ Angl. *goodness of fit of a distribution*.

skirstinį $F(x)$, kai n yra pakankamai didelis.

Kažkoku mums patogiu būdu apibrėžkime neneigiamą F_n^* nuokrypį nuo F ir pažymėkime jį raide D . Pavyzdžiui, D gali būti apibrėžta, taip:

$$D = \sup_x |F_n^*(x) - F(x)|.$$

Žinoma, nuokrypį D galima apibrėžti įvairiai, tačiau bet kuriuo atveju tai bus F_n^* funkcija. Todėl, žinodami empirinį skirstinį F_n^* , mes galime apskaičiuoti tikimybę $P(D > D_0)$, t.y. tikimybę įvykio, kad nuokrypis D viršys duotą fiksuotą teigiamą skaičių D_0 .

4. Šią tikimybę galima padaryti labai maža, jei parinkti pakankamai didelį D_0 . Parinkime jį taip, kad būtų teisingas sąryšis

$$P(D > D_0) = \alpha,$$

kur α yra pakankamai mažas skaičius. Žodžiai „pakankamai mažas“ čia reiškia tai, kad, kaip buvo minėta šio skyrelio pradžioje, mes esame praktiškai tikri, kad, atliekant eksperimentą tik vieną kartą, nagrinėjamas įvykis neįvyks.

5. Tarkime dabar, kad yra duota imtis, kurios dydis yra n . Pagal šios imties reikšmes apskaičiuojame nuokrypį D . Jei išaiškėja, kad $D > D_0$, tai reiškia, kad įvykis, kurio tikimybė yra α , įvyko. Kadangi α – pakankamai mažas skaičius, tai darome išvadą, kad atlikus eksperimentą, mūsų hipotezė yra atmetama.

Iš kitos pusės, jei, atlikus eksperimentą ir apskaičiavus nuokrypį D išaiškėja, kad $D \leq D_0$, tai mes priimame hipotezę ir skaitome, kad ji pakankamai tiksliai atspindi turimus duomenis.

6. Ši samprotavimo schema labai dažnai sutinkama, atliekant statistines išvadas ir formuluojant bei pagrindžiant statistinius kriterijus.

Pagal Lietuvos standartą, **statistinis testas**¹, **statistinis kriterijus** – tai statistinė procedūra, pagal kurią sprendžiama, ar reikia nulinę hipotezę² atmesti alternatyviosios hipotezės naudai, ar nereikia.

Standarte [8] pateikiami tokie pavyzdžiai:

a) Testas (kriterijus) tikrinti hipotezei, kad atsitiktinio dydžio X vidurkis μ yra ne mažesnis negu duotoji reikšmė μ_0 ;

¹ Angl. *statistical test*; rus. *статистический критерий*.

² Nulinės ir alternatyviosios hipotezių sąvokos išnagrinėtos 19-ame skyrelyje.

b) Testas (kriterijus) tikrinti hipotezei, kad defektingųjų elementų dalis vienoje partijoje p_1 ir defektingųjų elementų dalis kitoje partijoje p_2 turi tą pačią (nenurodytą) reikšmę;

c) Testas (kriterijus) tikrinti hipotezei, kad atsitiktinis dydis X yra pasiskirstęs normaliai (su nenurodytais parametrais).

7. Garsus švedų matematikas H. Krameras savo klasikinėje monografijoje [7] rašo taip:

„Paprasčiausiu atveju, kai išaiškėja empirinių reikšmių aibės skirstinio ir teorinio skirstinio atitikimas, mes naudojame specialų terminą ir kalbame apie *suderinamumo kriterijų*¹. Tikimybę α , kurią galime pasirinkti laisvai, vadiname *kriterijaus reikšmingumo lygmeniu*².

Tuo atveju, kai mūsų pasirinktas nuokrypis D viršija *reikšmingumo ribą* D_0 , mes skaitysimė, kad hipotezė atmesta, atlikus eksperimentą (praktiškai atmesta). Tačiau toks atmetimas, žinoma, nėra ekvivalentus loginiam atmetimui. Juk net jei hipotezė teisinga, įvykis $D > D_0$, turintis tikimybę α , atskiru išimtinu atveju visgi gali įvykti. Tačiau jei α pakankamai mažas skaičius, tai mes turime teisę praktikoje atmesti (nenagrinėti) tokią galimybę“.

8. Taigi, jei mums svarbu, kad eksperimento rezultatas būtų patikimas, α turi būti mažas, t.y. jei mes norime, kad nulinė hipotezė H_0 nebūtų atmesta nepagrįstai, reikšmingumo lygmuo turi būti mažas.

Pagal Lietuvos standartą, aibė kriterijaus statistikos reikšmių, kurioms nulinė hipotezė yra atmetama, yra vadinama *kritine sritimi*³.

Kritinės sritys apibrėžiamos taip, kad jei nulinė hipotezė yra teisinga, tai jos atmetimo tikimybė turi būti ne didesnė negu duotoji α reikšmė.

9. Tame pačiame standarte pateikiamas toks pavyzdys, kurį mes faktiškai išnagrinėjome 20-ame skyrelyje ir apibendrinome 21-o skyrelio 1-ame punkte. Klasikinis būdas tikrinti hipotezei $H_0 : \mu \geq \mu_0$ apie normaliojo skirstinio vidurkį, kai yra žinomas standartinis nuokrypis σ , su alternatyva $H_1 : \mu < \mu_0$, pagrįstas imties empirinio vidurkio statistika $\bar{X}$.

¹ Apie suderinamumo kriterijus detaliau kalbame sekančiuose skyreliuose.

² Angl. *significance level* (of a test); rus. *уровень значимости* (критерия).

³ Angl. *critical region*. 20-ame skyrelyje matėme, kad šis terminas turi sinonimą – *atmetimo sritis* (rejection region).

Čia kritinė sritis yra aibė reikšmių, mažesnių už

$$D_0 = \mu u_0 - u_{1-\alpha} \sigma / \sqrt{n},$$

kur n yra imties dydis, $u_{1-\alpha}$ – standartinio normaliojo atsitiktinio dydžio $(1 - \alpha)$ -kvantilis.

Šiuo atveju statistinis kriterijus yra toks: jei apskaičiuotoji statistikos $\bar{X}$ reikšmė $\bar{x}$ yra mažesnė už D_0 , tai H_0 atmetama. Priešingu atveju H_0 neatmetama (priimama).

43. CHI KVADRATO KRITERIJUS (K. PIRSONO SUDERINAMUMO TESTAS)

1. Uždavinys, kurį nagrinėsime šiame skyrelyje, yra *skirstinio atitikimo uždavinys*. Turime populiacijos imtį, kurios dydis yra n . Norime sužinoti, ar šios populiacijos skirstinys tikrai sutampa su teoriniu skirstiniu $F(x)$, kurį mes pasirinkome.

2. Norėdami išspręsti šį uždavinį, nuline hipoteze H_0 pasirinkime hipotezę apie tai, kad mūsų duomenys sudaro didumo n imtį, paimtą iš populiacijos, kurios skirstinys yra $F(x)$. Tarsime taip pat, kad $F(x)$ yra pilnai apibrėžtas, t.y. visi jo parametrai yra žinomi. Tai reiškia, kad H_0 yra paprastoji hipotezė. Patikslinsime suformuluotą uždavinį: *pateikti metodą (kriterijų, testą), leidžiantį patikrinti, ar mūsų duomenys yra suderinami su hipoteze H_0 .*

3. Pagal Glivenko teorema, empirinis skirstinys $F_n^*(x)$ gerai aproksimuoja populiacijos skirstinį $F(x)$, jei n yra pakankamai didelis ir jei hipotezė H_0 yra teisinga. Praeitame skyrelyje minėjome, kad natūralu apibrėžti neneigiamą F^* nuokrypį nuo F . Įvedus tokį matą, natūralu grįsti savo kriterijų (testą) šio mato empirinio skirstinio savybėmis.

Tokius nuokrypių matus galima konstruoti įvairiai, bet labiausiai paplitęs yra matas, susijęs su chi kvadrato kriterijumi¹, kurį prieš šimtą metų įvedė K. Pirsonas (K. Pearson).

Tarkime, kad stebimo atsitiktinio dydžio teikšmių sritis yra padalinta į baigtinį skaičių r nesikertančių intervalų $(-\infty; x_1), [x_1; x_2), \dots, [x_{r-1}; +\infty)$. Tegu $p_1, \dots, p_r$ – atitinkamos skirstinio $F(x)$ reikšmės šiuose intervaluose, t.y.

$$p_i = F(x_i) - F(x_{i-1}), \quad i = 1, 2, \dots, r, \quad \text{kur } x_0 = -\infty, x_r = +\infty.$$

¹ Angl. *chi-squared test*; χ^2 test; rus. *критерий χ^2* .

Tarkime, kad visiems i $p_i > 0$ ir pažymėkime ν_i imties elementų skaičių intervale $[x_{i-1}; x_i)$. Akivaizdu, kad

$$\sum_{i=1}^r \nu_i = n.$$

Sekanti problema mūsų tyrimo kelyje formuluojama taip: *rasti tinkamą nuokrypį tarp empirinio skirstinio F_n^* ir hipotezinio populiacijos skirstinio F matą.*

5. Skirstinio F masė, sukoncentruota i -tajame intervale $[x_{i-1}; x_i)$, yra p_i , o empirinio skirstinio F_n^* masė, sukoncentruota šiame intervale, yra ν_i/n (prisiminkime santykinio dažnio sąvoką, išnagrinėtą 3-jame skyrelyje). Vadovaudamiesi principu, analogišku mažiausių kvadratų metodui (žr. 24-ąjį skyrelį), *nuokrypį tarp F_n^* ir F matu* pasirinkime sumą

$$\sum_{i=1}^r c_i \left(\frac{\nu_i}{n} - p_i \right)^2, \quad (1)$$

kur teigiami koeficientai c_i gali būti pasirinkti laisvai.

K. Pirsonas parodė, kad, pasirinkus $c_i = n/p_i$, matas (1) įgija labai naudingas savybes, o patį matą (1) pažymėjo $\chi^2(n)$:

$$\chi^2(n) = \sum_{i=1}^r \frac{n}{p_i} \left(\frac{\nu_i}{n} - p_i \right)^2 = \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i}. \quad (2)$$

6. Naudingiausia savybė yra ta, kad augant imties dydžiui n , dydžio $\chi^2(n)$ empirinis skirstinys konverguoja į skirstinį, visiškai nepriklausantį nuo hipotezinio populiacijos skirstinio F .

K. Pirsono teorema. *Atsitiktinio dydžio $\chi^2(n)$ ribinis skirstinys, kai $n \rightarrow \infty$, yra chi kvadrato skirstinys su $r - 1$ laisvės laipsniai.*

Prisiminsime (žr. 18-ąjį skyrelį), kad *chi kvadrato skirstinys* – tai nepriklausomų normaliųjų standartinių atsitiktinių dydžių kvadratų sumos skirstinys.

Teoremos įrodymas detalčiau nagrinėjamas H. Kramero knygoje [7].

7. Dabar šios teoremos pagalba nesunku suformuluoti kriterijų, leidžiantį patikrinti, ar mūsų duomenys yra suderinami su hipoteze H_0 .

Tegu $\chi^2_{r-1,\alpha}$ yra chi kvadrato skirstinio su $r - 1$ laisvės laipsniu lygio α kritinė reikšmė (žr. 18-ąjį skyrelį), t.y.

$$P(\chi^2 \geq \chi^2_{r-1,\alpha}) = \alpha,$$

kur χ^2 yra atsitiktinis dydis, turintis chi kvadrato skirstinį su $r - 1$ laisvės laipsniu.

Pagal K. Pirsono teoremą, tikimybė

$$P(\chi^2(n) \geq \chi^2_{r-1,\alpha}) = \alpha_n$$

yra apytiksliai lygi α , jei tik n yra pakankamai didelis.

Fiksuokime dabar pakankamai mažą α . Tai reiškia, kad praktiškai galima skaityti, jog atlikus vieną eksperimentą, įvykis $\{\chi^2 \geq \chi_\alpha\}$ su tikimybe α neįvyks. Jei imties dydis n yra pakankamai didelis, tai praktiniams tikslams tikimybę α_n galima sutapatinti su jos ribine reikšme α . Taigi, *jei hipotezė H_0 yra teisinga, tai atliktame eksperimente praktiškai neįmanoma gauti $\chi^2(n)$ reikšmę, viršijančią χ^2_α .*

8. Pats kriterijus, o tuo pačiu ir atsakymas į punkte 2 suformuluotą uždavinį, yra toks.

Jei atlikus konkretų eksperimentą, gauname kad $\chi^2(n) \geq \chi^2_{r-1,\alpha}$, t.y. statistika $\chi^2(n)$ įgijo reikšmę iš *kritinės srities* (žr. 20-ąjį skyrelį), tai sakoma, kad yra *reikšmingas nuokrypis* nuo hipotezės H_0 ir mes turime šią hipotezę atmesti.

Priminsime, kad statistika, kurios pagrindu formuluojama sprendimo priimimo taisyklė, vadinama *kriterijaus statistika* (žr. 20-ąjį skyrelį). Taigi, galime tvirtinti, kad $\chi^2(n)$ yra *Pirsono kriterijaus statistika*.

O pagal Lietuvos standartą, **reikšmingas rezultatas**¹ (su pasirinktu reikšmingumo lygmeniu α) – tai statistinio kriterijaus (testo) rezultatas, kai nulinė hipotezė H_0 atmetama. Jei hipotezė neatmetama, tai rezultatas nėra reikšmingas.

Tikimybė, kad šioje situacijoje (t.y. kai $\chi^2(n) \geq \chi^2_{r-1,\alpha}$) nulinė hipotezė H_0 yra atmetama, nors iš tikrųjų yra teisinga, tiksliai lygi α_n ir apytiksliai lygi α . Štai kodėl tokiu atveju sakoma, kad *kriterijaus reikšmingumo lygmuo* yra α .

¹ Angl. *significant result*.

Pagal Lietuvos standartą, klaida, padaroma atmetant nulinę hipotezę (nes statistika įgijo reikšmę iš kritinės srities), kai iš tikrųjų nulinė hipotezė yra teisinga, vadinama **pirmosios rūšies klaida**. Ji dažnai vadinama *I tipo klaida*.

Pirmosios rūšies klaidos tikimybė vadinama tikimybė padaryti pirmosios rūšies klaidą. Kartais ji vadinama *I tipo rizika*.

Skaičius, nurodantis *I rūšies klaidos tikimybės viršutinį režį*, vadinamas **kriterijaus (testo) reikšmingumo lygmeniu**¹. Reikšmingumo lygmuo paprastai yra žymimas α .

9. Suformulavę pagal standartą eilę sąvokų, tęsime pradėtą statistinio kriterijaus nagrinėjimą.

Jei, atlikus konkretų eksperimentą, gauname, kad $\chi^2(n) < \chi^2_{r-1, \alpha}$, tai stebėjimų duomenys nulinei hipotezei H_0 neprieštarauja, šie duomenys yra *suderinami* su hipoteze H_0 . Žinoma, net tokiu konkrečiu atveju negalima teigti, kad hipotezė H_0 tikrai yra teisinga.

Hipotezės H_0 priėmimas byloja tik tai, kad ji neprieštarauja turimiems stebėjimų duomenims, yra su jais *suderinama*. Eksperimento išplėtimas ir naujų stebėjimo duomenų surinkimas pateikia medžiagą naujam hipotezės patikrinimui. Tai atitinka bendrus principus, pavyzdžiui, gamtos moksluose, kur bet kokia hipotezė gali būti atmesta gavus naujus, prieštaraujančius jai stebėjimų duomenis.

Pagal Lietuvos standartą, klaida, padaroma neatmetant (priimant) nulinę hipotezę (nes statistikos reikšmė nepateko į kritinę sritį), kai iš tikrųjų nulinė hipotezė yra neteisinga, vadinama **antrosios rūšies klaida**. Ji dažnai vadinama *II tipo klaida*.

Antrosios rūšies klaidos tikimybė vadinama tikimybė padaryti antrosios rūšies klaidą. Jos reikšmė, kuri paprastai žymima β , priklauso nuo realios situacijos ir gali būti apskaičiuota tik tada, kai atitinkamai konkretizuota alternatyvioji hipotezė. Kartais ji vadinama *II tipo rizika*.

10. Skyrelio pabaigoje kompaktiškai suformuluosime chi kvadrato kriterijų, kuris kartais vadinamas *Pirsono suderinamumo testu* arba tiesiog *suderinamumo testu*². Matematikos terminų žodynas rekomenduoja dar vieną terminą – *suderintumo kriterijus*.

Taigi, turėdami stebėjimo duomenis (imtį didumo n) ir lygindami $\chi^2(n)$,

¹ Angl. *significance level of a test*.

² Angl. *goodness-of-fit test*.

gautą šių duomenų pagrindu pagal formulę (2), su lygio α kritine reikšme $\chi^2_{r-1, \alpha}$, kurią apskaičiuojame lentelių pagalba, priimame vieną iš dviejų galimų sprendimų:

(A) Jei $\chi^2(n) \geq \chi^2_{r-1, \alpha}$, tai nulinė hipotezė H_0 yra atmetama alternatyviosios hipotezės naudai.

(B) Jei $\chi^2(n) < \chi^2_{r-1, \alpha}$, tai skaitoma, kad nėra pagrindo atmesti nulinę hipotezę, t.y. hipotezinis skirstinio pavidalas $F(x)$ neprieštarauja praktinių stebėjimų duomenims.

44. CHI KVADRATO KRITERIJAUŠ PRAKTINIS TAIKYMAS

1. Chi kvadrato kriterijaus praktinio taikymo nagrinėjimą pradėsime konkrečiu pavyzdžiu, o po to praeitame skyrelyje išnagrinėtos teorinės medžiagos pagrindu atliksime kai kuriuos apibendrinimus.

2. **Pirmas pavyzdys**¹. Tikimasi, kad nagrinėjamoje populiacijoje žmonių su I, II, III ir IV kraujo grupėmis proporcija yra atitinkamai 0,3; 0,4; 0,2 ir 0,1. Atsitiktinai pasirinktus ir apklaustus 400 žmonių gauta, kad I kraujo grupę turi 106, II – 148, III – 96, IV – 50 žmonių. Esant 5 procentų reikšmingumo lygmeniui patikrinti, ar gauti duomenys yra suderinami su aukščiau pateikta hipoteze.

3. *Sprendimas.* Pažymėkime p_i tikimybę įvykio, kad asmuo turi i -tąją kraujo grupę. Tada

$$H_0 : P_1 = 0,3, \quad p_2 = 0,4, \quad p_3 = 0,2, \quad p_4 = 0,1.$$

Norėdami apskaičiuoti nuokrypio matą $\chi^2(n)$ reikšmę pagal formulę (2) iš praeito skyrelio, pirmiausia turime apskaičiuoti hipotezinį dažnį np_i kiekvienoje žmonių grupėje. Kadangi $n = 400$ ir $p_1 = 0,3$, tai hipotezinis žmonių skaičius su pirmąja kraujo grupe yra lygus $400 \cdot 0,3 = 120$. Analogiškai įsitikiname, kad $np_2 = 160$, $np_3 = 80$, $np_4 = 40$. Skaičiavimų, reikalingų nuokrypio matą $\chi^2(400)$ reikšmei rasti, rezultatai pavaizduoti 18 lentelėje.

Kadangi imties elementai yra padalinti į keturias grupes, tai chi kvadrato skirstinys turi $4 - 1 = 3$ laisvės laipsnius. Iš chi kvadrato skirstinio lentelių randame, kad $\chi^2_{3,0.05} = 7,815$.

¹ Paimtas iš R. Khazanie knygos [6].

18 lentelė

Kraujo grupė	Stebimas dažnis γ_i	Hipotezinis dažnis np_i	$\nu_i - np_i$	$(\nu_i - np_i)^2$	$(\nu_i - np_i)^2 / np_i$
I	106	120	-14	196	1,63
II	148	160	-12	144	0,90
III	96	80	16	256	3,20
IV	50	40	10	100	2,50
Viso	400	400	0	696	8,23

Iš kitos pusės, pagal 18 lentelę

$$\chi^2(n) = \chi^2(400) = \sum_{i=1}^4 (\nu_i - np_i)^2 / np_i = 8,23.$$

Kadangi apskaičiuotoji reikšmė 8,23 yra didesnė už teorinę 7,815, tai pagal chi kvadrato kriterijaus punktą A (žr. praeitą skyrelį, p. 10) nulinė hipotezė H_0 yra atmetama.

4. Antras pavyzdys. Buvo tikrinamas 200 betono ruošinių atsparumas. Gauti tokie rezultatai:

19 lentelė

Atsparumo intervalai (kg/cm ²)[x_{i-1}, x_i)	Ruošinių skaičius
190–200	10
200–210	26
210–220	56
220–230	64
230–240	30
240–250	14

Apskaičiavus atsitiktinio dydžio X , išreiškiančio betono ruošinių atsparumo dydį, empirinį vidurkį $\bar{X}$ ir nepaslinktąją empirinę dispersiją S^2 gauta, kad $\bar{x} = 221$, o $s = 12,36$. Patikrinti nulinę hipotezę H_0 , suformuluotą šių duomenų pagrindu – $H_0 : X$ turi normalųjį skirstinį su vidurkiu $\mu = 221$ ir standartiniu nuokrypiu $\sigma = 12,36$.

Reikšmingumo lygmuo α yra 5%.

5. *Sprendimas.* Pirmiausia patikrinsime, kad $\bar{x} = 221$, o $s = 12,36$. Ištikrųjų, jei x_i^* – i -tojo atsparumo intervalo vidurio taškas, tai

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^6 x_i^* m_i = (195 \times 10 + 205 \times 26 + 215 \times 56 + 225 \times 64 \\ &\quad + 235 \times 30 + 245 \times 14) / 200 = 221; \\ s^2 &= \frac{1}{n-1} \sum_{i=1}^6 (x_i^* - \bar{x})^2 = ((-26)^2 \times 10 + (-16)^2 \times 26 \\ &\quad + (-6)^2 \times 56 + 4^2 \times 64 + 14^2 \times 30 + 24^2 \times 14) / 199 = 152,767; \\ s &= \sqrt{152,767} = 12,36.\end{aligned}$$

Toliau apskaičiuosime p_i – hipotezinio skirstinio masę, sukoncentruotą i -tajame intervale $[x_{i-1}, x_i]$. Jei $N_{\mu, \sigma}$ – normalusis atsitiktinis dydis su vidurkiu μ , standartiniu nuokrypiu σ ir, be to, $u_i = (x_i - \mu) / \sigma \geq 0$ ($i = 1, 2, \dots, 6$), tai

$$\begin{aligned}p_i &= P(x_{i-1} \leq N_{\mu, \sigma} < x_i) = P\left(\frac{x_{i-1} - \mu}{\sigma} \leq \frac{N_{\mu, \sigma} - \mu}{\sigma} < \frac{x_i - \mu}{\sigma}\right) \\ &= P\left(\frac{x_{i-1} - \mu}{\sigma} \leq Z < \frac{x_i - \mu}{\sigma}\right) \\ &= P(u_{i-1} \leq Z < u_i) = \int_0^{u_i} p(x) dx - \int_0^{u_{i-1}} p(x) dx.\end{aligned}\quad (1)$$

Čia $p(x)$ – standartinio normalaus atsitiktinio dydžio Z tankis.

Kadangi $p(x)$ – simetrinė funkcija, tai tais atvejais, kai bent vienas iš u_i nėra teigiamas, teisingi sąryšiai:

$$p_i = \int_0^{|u_{i-1}|} p(x) dx - \int_0^{|u_i|} p(x) dx, \quad \text{jei } u_i \leq 0, u_{i-1} \leq 0; \quad (2)$$

$$p_i = \int_0^{u_i} p(x) dx + \int_0^{|u_{i-1}|} p(x) dx, \quad \text{jei } u_{i-1} \leq 0 \leq u_i. \quad (3)$$

Dabar jau lengva pagal normaliojo skirstinio lenteles (Priedas, 1 lentelė) apskaičiuoti visus p_i , $i = 1, 2, 3, 4, 5, 6$, prieš tai, žinoma, apskaičiavus visus u_i

(taupydami vietą, pateikiame tik pirmos eilutės skaičiavimo rezultatus):

$$u_0 = (x_0 - \mu)/\sigma = (190 - 221)/12,36 = -2,51;$$

$$u_1 = (x_1 - \mu)/\sigma = (200 - 221)/12,36 = -1,70;$$

$$p_1 = \int_0^{|u_0|} p(x) dx - \int_0^{|u_1|} p(x) dx = 0,4940 - 0,4554 = 0,0386;$$

$$np_1 = 200 \cdot 0,0386 = 7,72; \quad \nu_1 - np_1 = 10 - 7,72 = 2,28;$$

$$(\nu_1 - np_1)^2 = 5,20; \quad (\nu_1 - np_1)^2 / np_1 = 0,67.$$

Sudarome lentelę:

20 lentelė

Stebimo dydžio X reikšmių intervaliniai pokyčiai $[x_{i-1}; x_i)$	Stebimas dažnis ν_i	Normuoti ir centruoti intervalai $[u_{i-1}; u_i)$	Hipotezinio skirstinio masė p_i	Hipotezinis dažnis np_i	$(\nu_i - np_i)^2$	$(\nu_i - np_i)^2 / np_i$
190–200	10	$[-2,51; -1,70)$	0,0386	7,72	5,20	0,67
200–210	26	$[-1,70; -0,89)$	0,1421	28,42	5,86	0,21
210–220	56	$[-0,89; -0,08)$	0,2814	56,28	0,08	0,00
220–230	64	$[-0,08; 0,73)$	0,2992	59,84	17,31	0,29
230–240	30	$[0,73; 1,54)$	0,1709	34,18	17,47	0,51
240–250	14	$[1,54; 2,35)$	0,0524	10,48	12,39	1,18
Viso:	200	$[-2,51; 2,35)$	0,9846	196,92	58,31	2,86

Atkreipsime dėmesį į tai, kad pirmose trijose lentelės eilutėse ir u_{i-1} , ir u_i yra neigiami, todėl skaičiuojant p_i taikoma formulė (2). Ketvirtoje eilutėje $u_3 < 0$, bet $u_4 > 0$, todėl taikoma formulė (3). Penktoje ir šeštoje eilutėse $u_{i-1} > 0$, $u_i > 0$, todėl skaičiuojant p_5 ir p_6 taikoma formulė (1).

Visa hipotezinio skirstinio masė, suprantama, lygi 1. Kodėl tada gavome, kad

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 0,9846?$$

Tai galima paaiškinti tuo, kad vietoj teorijoje nagrinėjamos hipotezinio skirstinio apibrėžimo srities $(-\infty, \infty)$ praktikoje mes apsiribojome intervalu $[-2,51; 2,35)$. Be to, reikia neužmiršti apvalinimo paklaidos.

Iš lentelės matome, kad

$$\chi^2(n) = \chi^2(200) = \sum_{i=1}^6 (\nu_i - np_i)^2 / np_i = 2,86.$$

Kadangi imties elementai yra padalinti į šešias grupes, tai chi kvadrato skirstinys turi $6 - 1 = 5$ laisvės laipsnius. Iš chi kvadrato skirstinio lentelių (Priedas, 3 lentelė) randame, kad

$$P(\chi^2 \geq \chi_{5,0.05}^2) = P(\chi^2 \geq 11,070) = 0,05;$$

t.y. $\chi_{5,0.05}^2 = 11,070 > 2,86 = \chi^2(200)$.

Pagal chi kvadrato kriterijaus punktą B (žr. praeitą skyrelį, p. 10) nėra pagrindo atmesti nulinę hipotezę H_0 apie atsitiktinio dydžio X normalumą su vidurkiu 221 ir standartiniu nuokrypiu 12,36.

6. Apibendrinant išnagrinėtus pavyzdžius, galima pateikti tokią *chi kvadrato kriterijaus praktinio taikymo schemą*:

1⁰. Stochastinio eksperimento rezultatų reikšmių diapazoną padaliname į r vienodo ilgio intervalų.

2⁰. Suskaičiuojame kiek i -tajame intervale yra nagrinėjamos imties, kurios dydis yra n , elementų. Jų skaičių pažymime ν_i .

3⁰. Hipotezinio skirstinio masę i -tajame intervale pažymime p_i ir apskaičiuojame ją pagal lenteles ar kitokiu būdu.

4⁰. Apskaičiuojame hipotezinį dažnį np_i .

5⁰. Hipotezinio ir stebimo skirstinių nuokrypio matas $\chi^2(n)$ apibrėžiamas formule

$$\chi^2(n) = \sum_{i=1}^r (\nu_i - np_i)^2 / np_i.$$

ir apskaičiuojamas pagal 21 lentelę.

6⁰. Jei nulinė hipotezė H_0 yra teisinga, n – pakankamai didelis ir nei vienas iš hipotezinių dažnių $np_1, \dots, np_r$ nėra mažesnis už 5, tai atsitiktinio dydžio $\chi^2(n)$ skirstinys yra pakankamai artimas chi kvadrato skirstiniui su $r - 1$ laisvės laipsniais. Ši išvada yra pagrindas *sprendimo priėmimo taisyklės* formulavimui.

7⁰. **Sprendimo priėmimo taisyklė** yra tokia:

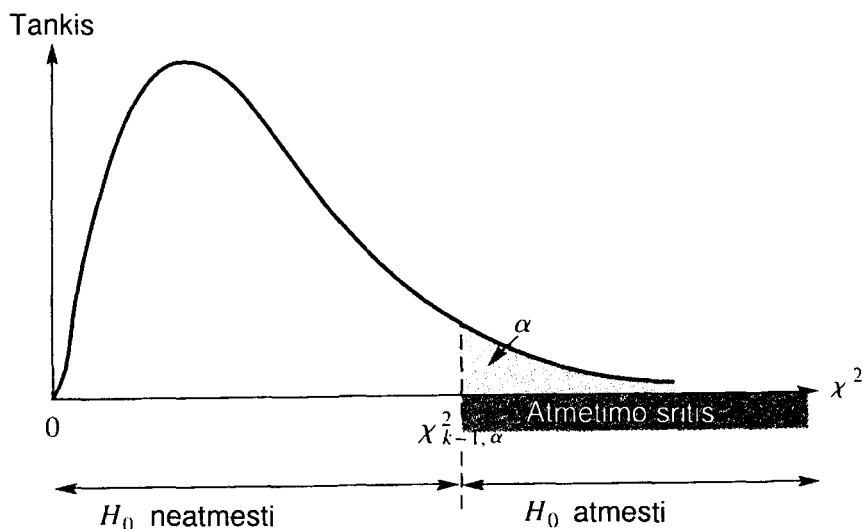
H_0 atmesti, jei pagal 21 lentelę apskaičiuotoji $\chi^2(n)$ reikšmė yra didesnė už $\chi_{r-1,\alpha}^2$ reikšmę, apskaičiuotą pagal χ^2 skirstinio lentelę (Priedas, 3 lentelė).

21 lentelė

Stebimas intervalas $[x_{i-1}, x_i)$	Stebimas dažnis ν_i (imties elementų skaicius in-le $[x_{i-1}, x_i)$	Hipotezinis dažnis np_i	Nuokrypio mato dėmuo $(\nu_i - np_i)^2 / np_i$
$[x_0, x_1)$	ν_1	np_1	$(\nu_1 - np_1)^2 / np_1$
$[x_1, x_2)$	ν_2	np_2	$(\nu_2 - np_2)^2 / np_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_{r-1}, x_r)$	ν_r	np_r	$(\nu_r - np_r)^2 / np_r$
Viso	n	n	$\chi^2(n)$

Čia α – reikšmingumo lygmuo.

Grafiškai sprendimo priėmimo taisyklę galima pavaizduoti taip:



46 pav. Atmetimo (kritinė) sritis, kai kriterijaus statistika pasirenkamas χ^2 skirsny

45. KOLMOGOROVO IR SMIRNOVO KRITERIJAI SUDERINAMUMO IR VIENODO PASISKIRSTYMO HIPOTEZĖMS TIKRINTI

1. Nagrinėdami *skirstinio atitikimo* (žr. 42-ąją skyrelį) problemą, neneigiamą empirinio skirstinio F_n^* nuokrypio matą nuo hipotetinio skirstinio F pažymėjome raide D ir sakėme, kad šis nuokrypis apibrėžiamas „kažkokiu mums

patogiū būdu“. K. Pirsonas, kaip mes matėme 43-jame skyrelyje, nagrinėdami chi kvadrato kriterijų, nuokrypio tarp F_n^* ir F matu pasirinko sumą

$$D = \sum_{i=1}^r c_i \left(\frac{\nu_i}{n} - p_i \right)^2.$$

A. Kolmogorovas pasiūlė nuokrypio matu D (o tuo pačiu ir *kriterijaus statistika*) pasirinkti funkciją D_n ,

$$D_n = \sup_{-\infty < x < \infty} |F_n^*(x) - F(x)|.$$

2. Šis A. Kolmogorovo nuokrypio mato pasirinkimas buvo grindžiamas tokia teorema.

A. Kolmogorovo teorema. *Jei teorinis skirstinys $F(x)$ yra tolydi funkcija, o $F_n^*(x)$ yra empirinis skirstinys, tai*

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n < x) = K(x) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\{-2k^2 x^2\} \quad (1)$$

bet kuriame fiksuotame taške $x > 0$.

Yra žinoma, kad ribinis skirstinys K (kuris yra vadinamas *Kolmogorovo skirstiniu*) pakankamai gerai aproksimuojamas jau kai $n = 20$. Tai reiškia, kad Kolmogorovo teoremos pagrindu pasirenkant *sprendimo priėmimo taisyklę* (žr. 20-tąjį skyrelį), praktiniuose skaičiavimuose pakanka pasirinkti imties dydį $n \geq 20$. Tokio tipo testas yra vadinamas *Kolmogorovo kriterijumi* suderinamumo hipotezėms tikrinti.

3. Funkcija $K(x)$ yra tabuluota, t.y. yra žinomos ir daugelyje statistikos knygų skelbiamos jos reikšmių lentelės. Pabrėšime, kad statistikos (nuokrypio mato) D_n ribinis skirstinys nepriklauso nuo F .

Palyginsime Kolmogorovo teoremos pagrindu konstruojamą kriterijų su pračiuose skyreliuose išnagrinėtu chi kvadrato kriterijumi.

- Chi kvadrato kriterijus formuluojamas bet kuriam teoriniam skirstiniui $F(x)$, o Kolmogorovo – tik tolygiajam;
- Kolmogorovo kriterijus tiksliau už chi kvadrato kriterijų atspindi hipotezės suderinamumą su empiriniais duomenimis.

- Chi kvadrato kriterijuje lyginami empiriniai ir teoriniai skirstinių dažniai, o taikant Kolmogorovo kriterijų lyginami empiriniai ir teoriniai skirstiniai.

4. Kolmogorovo teorema ypač patogiu pasinaudoti tada, kai reikia apskaičiuoti abipusius režius, kuriuose teorinis skirstinys $F(x)$ yra su duota tikimybe α .

Iš tikrųjų, tegu $t_{1-\alpha}$ yra skirstinio $\mathcal{K}(x)$ $1 - \alpha$ -kvantilis, t.y. $\mathcal{K}(t_{1-\alpha}) = 1 - \alpha$. Tada iš (1) formulės gauname, kad

$$\mathbf{P}(\sqrt{n}D_n \leq t_{1-\alpha}) = \mathbf{P}(F_n^*(x) - t_{1-\alpha}/\sqrt{n} \leq F(x) \leq F_n^*(x) + t_{1-\alpha}/\sqrt{n} \quad \forall x \in (-\infty, \infty)) \xrightarrow{n \rightarrow \infty} \mathcal{K}(t_{1-\alpha}) = 1 - \alpha.$$

Tai reiškia, kad, esant pakankamai dideliems n , su tikimybe artima skaičiui $1 - \alpha$, skirstinio $F(x)$ reikšmės visiems x tenkina nelygybes

$$F_n^*(x) - t_{1-\alpha}/\sqrt{n} \leq F(x) \leq F_n^*(x) + t_{1-\alpha}/\sqrt{n}.$$

Kadangi $0 \leq F(x) \leq 1$, tai šias nelygybes galima patikslinti:

$$\max(0, F_n^*(x) - t_{1-\alpha}/\sqrt{n}) \leq F(x) \leq \min(F_n^*(x) + t_{1-\alpha}/\sqrt{n}, 1).$$

Taip apibrėžta sritis vadinama *asimptotiniu dvipusiu pasiklivimo*¹ *intervalu* (su *pasiklivimo lygmeniu* $1 - \alpha$) teoriniam skirstiniui $F(x)$.

5. Suformuluosime dar vieną svarbų rezultatą, priklausančią N. Smirnovui ir gautą 1944 metais, kuris atskleidžia kitas naudingas empirinių skirstinių savybes.

Tegu yra duotos dvi atsitiktinės imtys, kurių dydžiai yra n_1 ir n_2 . Tegu šias imtis atitinkančios populiacijos turi (nežinomus) skirstinius $F_1(x)$ ir $F_2(x)$. Šių imčių pagrindu reikia patikrinti tokią nulinę hipotezę $H_0 : F_1(x) = F_2(x)$.

Ši hipotezė tikrinama *Smirnovo kriterijaus statistikos* (kai kurie autoriai vadina ją *Kolmogorovo-Smirnovo statistika*) analizės pagrindu, kuri tuo pačiu yra nuokrypio matas D_n^* :

$$D_n^* = \max_x |F_{1,n_1}^*(x) - F_{2,n_2}^*(x)|,$$

¹ Terminologija pagal Lietuvos standartą. Matematikos terminų žodynas nurodo terminą *pasikliautinasis intervalas*, *pasikliautinoji sritis*.

kur $n = n_1 n_2 / (n_1 + n_2)$, o F_{1,n_1}^* ir F_{2,n_2}^* – pirmosios ir antrosios imties empiriniai skirstiniai.

6. Mes ne kartą naudojomes tuo faktu, kad empirinis skirstinys yra pakankamai geras teorinio skirstinio įvertinys ir, augant imties dydžiui, jie suartėja. Todėl tuo atveju, kai yra teisinga hipotezė H_0 , empiriniai skirstiniai F_{1,n_1} ir F_{2,n_2} konverguoja į tą patį (nežinomą) teorinį skirstinį. Tokiu atveju, esant pakankamai dideliems n_1 ir n_2 , nuokrypio matas D_n^* neturėtų žymiai nukrypti nuo nulio.

Iš čia gauname, kad jei statistika (nuokrypio matas) D_n^* įgija pernelyg dideles reikšmes, tai šis faktas byloja hipotezės H_0 nenaudai.

Dabar Smirnovo teoremos pagrindu konkretizuosime aptartą idėją.

7. *N. Smirnovo teorema* tvirtina, kad jei nulinė hipotezė H_0 yra teisinga, tai statistikų $\{D_n^* \sqrt{n}\}$ skirstinių seka konverguoja į Kolmogorovo skirstinį $\mathcal{K}(x)$.

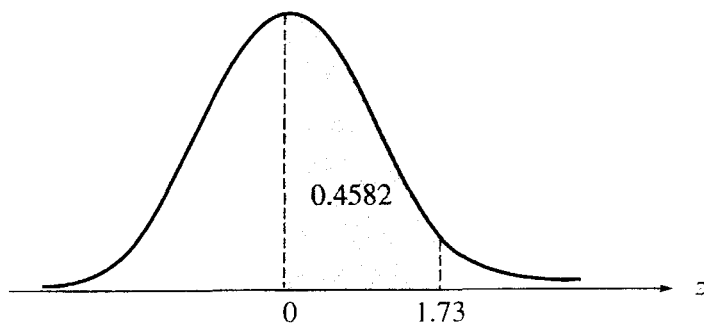
Smirnovo kriterijaus sprendimo priėmimo taisyklė yra analogiška chi kvadrato kriterijaus sprendimo priėmimo taisyklei. Iš tikrųjų, pirmiausia pasirenkame reikšmingumo lygmenį α . Tada iš sąryšio

$$\alpha = 1 - \mathcal{K}(t_\alpha) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 t_\alpha^2)$$

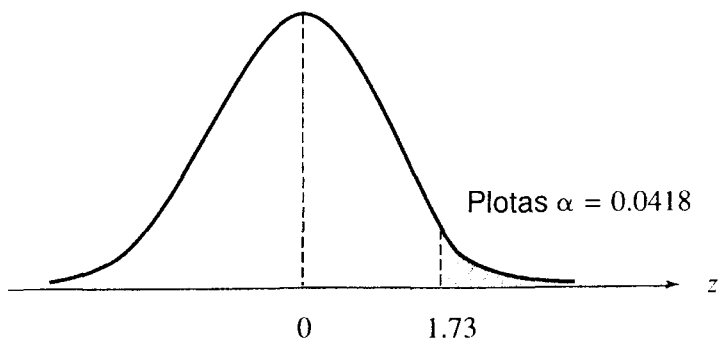
lentelių pagalba randame Kolmogorovo skirstinio α -kvantilį t_α . Dabar pagal turimas imčių reikšmes apskaičiuojame empirinės statistikos $D_n^* \sqrt{n}$ reikšmę, kurią pažymėsime d_{steb} ir palyginame d_{steb} su α -kvantiliu t_α .

Jei $d_{steb} \geq t_\alpha$, tai nulinė hipotezė H_0 yra atmetama. Priešingu atveju, t.y. kai $d_{steb} < t_\alpha$, yra skaitoma, kad nėra pagrindo atmeti hipotezę H_0 apie tai, jog abiejų nagrinėjamų imčių skirstiniai sutampa.

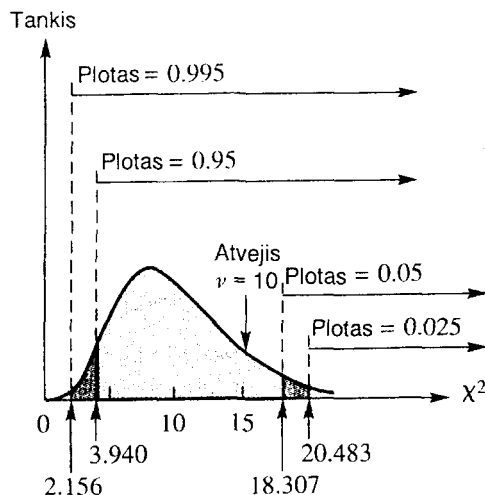
8. Smirnovo ar bet kuris kitas kriterijus vienodo pasiskirstymo hipotezėms tikrinti praktikoje taikomas tais atvejais, kai, pavyzdžiui, reikia nustatyti, kaip įrenginio daroma paklaida kinta nagrinėjamu laikotarpiu. Šie kriterijai taikomi ir tais atvejais, kai analizuojamos skirtingų produkcijos partijų kontrolinės imtys, norint patikrinti, kaip kito produkcijos kokybė, keičiantis pamainoms, ir pan.



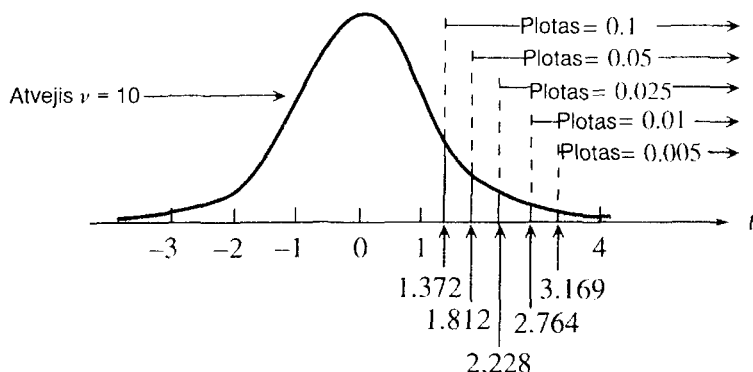
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4692	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010



Laisvės laipsnis ν	Dešiniojos uodegos po kreive plotas							✓
	0.995	0.99	0.975	0.95	0.05	0.025	0.01	
	Kritinės reikšmės							
	$\chi^2_{\nu, 0.995}$	$\chi^2_{\nu, 0.99}$	$\chi^2_{\nu, 0.975}$	$\chi^2_{\nu, 0.95}$	$\chi^2_{\nu, 0.05}$	$\chi^2_{\nu, 0.025}$	$\chi^2_{\nu, 0.01}$	$\chi^2_{\nu, 0.005}$
1	0.0 ¹ 393	0.0 ¹ 157	0.0 ¹ 982	0.0 ² 393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.652	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672



Laisvės laipsnis ν	α , dešimtosios uodegos po kreive plotas				
	0.1	0.05	0.025	0.01	0.005
	Kritinės reikšmės				
	$t_{\nu, 0.1}$	$t_{\nu, 0.05}$	$t_{\nu, 0.025}$	$t_{\nu, 0.01}$	$t_{\nu, 0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
...	...	...	...	...	...
∞	1.282	1.645	1.960	2.326	2.576

Naudota literatūra

1. **А.А. Боровков.** Математическая статистика. – Москва, Наука, 1984. – 472 с.
2. **И.И. Гихман, А.В. Скороход, М.И. Ядренко.** Теория вероятностей и математическая статистика. – Киев, Вища школа, 1979. – 408 с.
3. **K.J. Hastings.** Probability and statistics. – New York, Addison-Wesley, 1997. – 670 p.
4. **Г.И. Ивченко, Ю.И. Медведев.** Математическая статистика. – Москва, Высшая школа, 1984. – 248 с.
5. **R.A. Johnson, G.K. Bhattacharyya.** Statistics. Principles and methods (third edition). – New York, John Willey and Sons, 1996. – 720 p.
6. **R. Khazanie.** Statistics in a world of applications (fourth edition). – New York, HarperCollins, 1996. – 888 p.
7. **Г. Крамер.** Математические методы статистики. – Москва, Мир, 1975. – 648 с.
8. **Lietuvos standartas LST ISO 3534-1.** Statistika. Terminai ir apibrėžimai. – Vilnius, 1996. – 65 p.
9. **Lietuvos statistikos metraštis 1998.** – Vilnius, 1998. – 558 p.
10. **Matematikos terminų žodynas.** – Vilnius, Mokslo ir enciklopedijų leidykla, 1994. – 726 p.
11. **Р.В. Янушкявичюс.** Устойчивость характеристик вероятностных распределений. – Vilnius, Mokslo, 1991. – 248 p.

PAVEIKSLŲ IR JŲ KOMENTARŲ SĄRAŠAS

1. Susituokusių 1997 m. Lietuvos vyrų stulpelinė diagrama pagal amžių (psl. 17).
2. Susituokusių 1997 m. Lietuvos vyrų histograma pagal amžių (psl. 19).
3. Santykinio dažnio $n(H)/n$ svyravimų amplitudžių mažėjimas bei statistinio stabilumo fenomenas monetos mėtymo stochastiniame eksperimente (psl. 21).
4. Tikimybės masės funkcijos histogramos (tikimybines histogramas) pavyzdys (psl. 24).
5. Standartinio normaliojo atsitiktinio dydžio populiacijos stebinių modeliavimo rezultatai, iliustruojantys didžiųjų skaičių dėsnio veikimą (psl. 27).
6. Fizikinė vidurkio interpretacija – svertinės sąpuoklės (psl. 28).
7. Dvi stebinių aibės gali turėti tas pačias centro pozicijas, bet skirtingą sklaidos apie šį centrą pavidalą (psl. 30).
8. Tikimybinių tankio kreivė, pavaizduota kaip santykinio dažnio histogramų ribinė forma (psl. 34).
9. Plotas po atsitiktinio dydžio X tikimybinių tankio kreive tarp a ir b lygus tikimybei $P(a < X < b)$ (psl. 35).
10. Skirstinio reikšmė taške x lygi plotui po tikimybinių tankio kreive iki taško x (psl. 36).
11. Darbuotojo atvykimo į darbą santykinio dažnio histograma (psl. 37).
12. Tolygusis tikimybinių tankis kaip santykinių dažnių histogramų riba (psl. 38).
13. Konkrečios tikimybės apskaičiavimas pagal tolygiojo tikimybinių tankio grafiką (psl. 39).
14. Fizikinė vidurkio interpretacija tais atvejais, kai egzistuoja tankis (psl. 40).
15. Tankio su vienodais vidurkiu bei mediana pavyzdys (psl. 40).
16. Tankių su nevienodais vidurkiu bei mediana pavyzdžiai (psl. 40).
17. Cukraus fasavimo paklaidos histograma (psl. 42).
18. Kritulių skaičiaus per metus histograma (psl. 42).
19. Studentų žinių įvertinimų histograma (psl. 43).
20. Normaliojo tankio kreivės skirtingoms σ reikšmėms. Šios kreivės yra simetriškos μ atžvilgiu (psl. 43).
21. Standartinė normalioji kreivė. 95,4% ploto, esančio po šia kreive, sukoncentruota intervale $[-2;2]$ (psl. 44).

22. Normaliųjų tankių uodegos yra lygios (psl. 44).
23. Grafinis ir procentinis įvykių $\{\mu - k\sigma < N < \mu + k\sigma\}$ tikimybių vaizdavimas (psl. 45).
24. Centrinės ribinės teoremos praktinio „patikrinimo“ pavyzdyje tankis parinktas taip, kad jis gerokai skirtųsi nuo normaliojo tankio (psl. 55).
25. Centrinės ribinės teoremos „patikrinimo“ rezultatas – vidurkio $\bar{X}$ tankio histograma pakankamai gerai aproksimuojama normaliuoju tankiu (psl. 57).
26. Kiekviena standartinio normaliojo tankio uodega (kairėje ir dešinėje) lygi 0,025 (psl. 63).
27. Taškas $-z_{\alpha/2}$ yra standartinio normaliojo atsitiktinio dydžio Z $\alpha/2$ -kvantilis, kadangi $P(Z < -z_{\alpha/2}) = \alpha/2 = P(Z \leq -z_{\alpha/2})$ (psl. 64).
28. Pasiklivimo intervalų reikšmės. Jei μ būtų tokia, kaip pavaizduota, tai galėtume tvirtinti, kad ji nepriklauso tik trečiajam intervalui (psl. 66).
29. Didejant laisvės laipsniams, Stjudento tankis tampa vis labiau panašus į normalųjį tankį (psl. 68).
30. $t_{\nu, \alpha}$ yra Stjudento skirstinio su ν laisvės laipsnių lygmens α kritinė reikšmė (psl. 68).
31. Chi kvadrato tankių kreivės (psl. 71).
32. $\chi^2_{\nu, \alpha}$ yra chi kvadrato skirstinio su ν laisvės laipsnių lygmens α kritinė reikšmė (psl. 71).
33. Chi kvadrato skirstinio kritinių reikšmių radimas (psl. 72).
34. Nulinės hipotezės atmetimo sritis R (psl. 79).
35. Paprasčiausias dydžių tarpusavio priklausomybės atvejis – tiesinis sąryšis (psl. 83).
36. Tiesinis sąryšis, kuriame krypties koeficientas yra neigiamas (psl. 83).
37. Krypties koeficientas reprezentuoja y pokytį, x įgijus vienetinį prieauglį (psl. 84).
38. Sąryšio tarp studentų darbo rezultatų viso semestro metu ir egzamino metu tyrimo rezultatai (psl. 85).
39. (a)–(f) Koreliacijos koeficiento savybės (psl. 87–89).
40. Tiesės, kuri sklaidos diagramoje geriausiai atspindi tiesinį tendą, radimas (psl. 91).
41. Regresijos tiesė, gauta mažiausių kvadratų metodu (psl. 93).
42. Tolygiojo intervale $[a; b]$ skirstinio grafikas (psl. 110).

43. Dvitaškio skirstinio grafikas (psl. 110).
44. Imties, kurios dydis $n = 10$, empirinio skirstinio grafiko pavyzdys (psl. 119).
45. Silpnai konverguojančios į išsigimusį skirstinį $F_\infty(x)$ tolygiųjų skirstinių sekos $\{F_n(x)\}$ pavyzdys, kuri diverguoja tolygiojoje metrikoje (psl. 127).
46. Atmetimo (kritinė) sritis, kai kriterijaus statistika pasirenkamas χ^2 skirstinys (psl. 153).

A

Aibė:

- Borelio 100, 118
- elementariųjų baigčių 14, 94
- įvykių 13, 103
- neskaičiuojami 97
- skaičiuojami 97
- statistinių duomenų 10
- tuščių 14

Aksioma 103

Aksiomų sistema (Kolmogorovo) 103

– –, statistikos 103

Algebra:

- Borelio 100
- paprastoji 97
- σ -algebra 100
- , mažiausia 100

Aproksimacija normaliuoju skirstiniu 58

– geriausia 131

Asimptotinis normalumas 127, 129

Atranka 14

Atsitiktinis dydis 22, 105, 106

- – diskretusis 23, 32
- – dvitaškis 117
- – standartinis normalusis 47
- – tolydusis 22, 32, 36
- – tolygusis 36

Atsitiktinis intervalas 63

– įvykis 13, 15, 20, 103

– –, generuotas 107

Atstumas 133, 134

B

Baigtis elementarioji 14, 94

Balanso taškas 39

Bandymas 12

C, Č

Centrinė ribinė teorema 45, 55, 56, 126

Charakteristikos empirinės 121

Čebyševas 32

D

Dažnis 15, 16, 20

– kategorijos 15

– santykinis 11, 16, 20, 116

Dažnių lentelė 15

– skirstinys 16

Diagrama:

- medžio 12, 13
- sklaidos 85, 88
- stulpelinė 16

Didžiųjų skaičių dėsnis 26, 117

– – – sustiprintas 120, 122

Dispersija:

- atsitiktinio dydžio 32
- empirinė 31, 67, 121, 131
- imties 31, 67, 121
- populiacijos 67

Draudimo modelis 28

Duomenys:

- kiekybiniai 15
- kokybiniai 15
- pagal kategorijas 15
- skaitiniai 15
- suderinami 147

E, Ė

Eilutė variacinė 118

Eksperimentas:

stochastinis 10, 12

tikimybinis 10

Elementarioji baigtis 14, 94

Elementarusis įvykis 12, 14, 94

– –, palankusis 13

Elementas imčių erdvės 14

Empirinis (*kaip žodis*) 121

Empirinis matas 116

– skirstinys 118

Erdvė:

diskrečioji 94

elementariųjų įvykių 12, 94, 103

imčių 14

mačioji 102, 103

tikimybė 103

Ėmimas 49

– gražintinis 49, 51

– negražintinis 49, 52

Ėmimo vienetas 14

F

Fišeris R. 94

Funkcija:

Borelio 122

logaritminė tikėtinumo 137

mačioji 106

skirstinio 35, 108

tankio 33, 35

tikėtinumo 137

tikimybės masės 23

G

Grupė 16

H

Helmertas F. 70

Hipotezė 74, 77

– alternatyvioji 74

– neparametrinė 139, 140

– nepriklausomumo 139

– nulinė 74

– paprastoji 139

– parametrinė 139, 140

– suderinamumo 139

– sudėtinė 139

– vienodo pasiskirstymo 140

Hipotezės tikrinimo principas 81, 82

– – žingsniai 81

Hipotezių tipai 139

Histograma 18

– tikimybė 24

– santykinio dažnio 33

–, savybės 18

I, Į

Imties dydis 14

– paklaida 60

– dispersija 31, 67, 121

Imtis 14, 46

– konkrečioji 48, 61

Indikatorius 116

Interpretacija:

empirinė 65

statistinė 94

tikimybė 94

Intervalas:

atsitiktinis 63, 65

klasės 17

pasikliautinasis 155

pasiklivimo dvipusis 61, 155

- vienaspusis 62
- [vertinys 48, 59, 115, 124
- asimptotiškai normalus 127
- didžiausio tikėtinumo 136
- gautas momentų metodu 130
- geriausias 59, 115
- keitimo 125
- nepaslinktasis 61, 131, 132
- suderintasis 125
- – stipriai 125
- [vertis 48, 59, 61, 124
- [vykis:
 - atsitiktinis 13, 15, 20
 - būtinasis 14
 - elementarusis 12, 14, 94
 - palankus 13
 - negalimasis 14
- K
- Kazino 29
- Klaida:
 - pirmos rūšies 82, 147
 - antros rūšies 82, 147
- Klasė 16
- Klasės apatinė riba 18
- intervalas 17
- plotis 18
- viršutinė riba 18
- Koeficientas:
 - korekcijos 53
 - koreliacijos 86, 87
 - krypties 84, 85, 91
 - postūmio 84, 85, 91
 - regresijos 92
- Kolmogorovas A. 103, 154
- Konvergavimas:
 - pagal tikimybę 118
 - silpnas 126, 127
 - su tikimybe 120
 - tolygiojoje metrike 127
- Krameras H. 10, 58, 94, 143
- Kreivė:
 - normalioji 41, 42, 47
 - standartinė 44
 - tikimybinio tankio 33
- Kriterijaus reikšmingumo lygmuo 146, 147
- Kriterijaus statistika 79, 81, 154
- Kriterijų statistikos:
 - Kolmogorovo 154
 - Pirsono 146
 - Smirnov 155
- Kriterijus (testas):
 - chi kvadrato (Pirsono) 144, 147
 - –, taikymas 148, 152, 153
 - didelių imčių normalumo 80
 - Kolmogorovo 154
 - statistinis 141, 142
 - suderinamumo 143, 147
 - suderintumo 147
- Kritinė reikšmė 68, 70, 78, 146
- Kulback-Leibler atstumas 136
- Kvantis 41, 64
- Kvartilis 41
- L
- Ležandras A.M. 90
- Lygmuo:
 - kritinės reikšmės 68, 70, 146
 - kriterijaus reikšmingumo 146, 147
 - pasiklivimo 62, 155
 - reikšmingumo 79, 80

Lygtis:

didžiausio tikėtinumo 137

Koši 113

prognozės 90

tiesinio sąryšio 83, 85

M

Matas 103

– dominuojantis 133

– empirinis 116

– kitimo 30, 31

– Lebegio 105

– normuotas 20, 103

– nuokrypio 144, 145, 154, 155

– skaičiuojantysis 133

– sklaidos 30, 31

– tikimybinis 20, 95, 102, 103

Matematinė viltis 26, 53

Mediana 39, 41

Metodas:

keitimo 124, 125, 129

mažiausių kvadratų 90

momentų 130

pasiklivimo intervalo 61

Metrika:

stipresnė 127

tolygioji 127

Minimizacijos lema 134

Modelis diskretusis tikimybinis 95

– draudimo 28

Momentas:

empirinis 121

empirinis centruotas 121

– –, konvergavimas 124

N

Netiktis 90

Nebendramačiai skaičiai 113

Nuokrypis 31, 142

– reikšmingas 146

P

Paklaida:

imties 60

įvertinio 132

maksimali 64

prognozės 90

sistemoji 131

standartinė 54, 132

Paklaidos rešis 64

– – apytikslis 64

– – asimptotinis 64

Parametras 46

Pasiklivimo intervalas 61, 62, 64,

66, 69, 73

– lygmuo 62

Pirsonas K. 70, 144, 145, 147, 154

Plotis (klasės) 18

Populiacija 14

Poslinkis 131

– įvertinio 132

Požymis 16

Procentilis 41

R

Regresijos tiesė 91, 92, 93

Regresinė analizė 89

Reikšmė:

įvertinio 48

kritinė atmetimo 78

– chi kvadrato 32 pav.

– Stjudento 30 pav.

vidurkio 48

Reikšmingas rezultatas 146

Riba:

klasės apatinė 17

klasės viršutinė 17

reikšmingumo 143

Rinkimas 49

Rizika (I ir II tipo) 147

S

Savybė:

atminties neturėjimo 112

veiksmo be praeties poveikio 112

Sąryšis:

atvirkštinis 86

neigiamas 86, 88

teigiamas 85, 87

tiesinis 83

Skirstinio atitikimas 141, 143, 144

– funkcija 35, 108

Skirstinys 35, 108, 111

– chi kvadrato 70, 71, 145

– dažnių 16

– dvitaškis 109

– eksponentinis 56, 112

– empirinio vidurkio 50, 55

– empirinis 118

– gama 138

– Helmerto-Pirsono 70

– hipotezinis 139

– Kolmogorovo 154

– normalusis 41, 133

– Puasono 133

– standartinis normalusis 43, 133

– Stjudento 67

– tikimybių 108

– tolydusis 36

– tolygusis 36, 109

– vektoriaus 140

Smirnovas N. 155

Sritis:

– atmetimo 78, 143, 153 (46 pav.)

– kritinė 78, 143

– pasikliautinoji 155

Stabilūs rezultatai 11

Standartinis nuokrypis:

– atsitiktinio dydžio 32

– empirinis 31

– populiacijos 32

Standartizacija 44

Statistika (*kaip funkcija*) 46, 59, 81, 115, 124

– Kolmogorovo-Smirnovo 155

Statistika (*kaip mokslas*) 9, 10

Statistikos aksiomų sistema 103

– objektas 10

Statistinės išvados 45

Statistinio stabilumo fenomenas 11, 20, 33, 46

Statistinių duomenų aibė 10

Stebinys 14, 15, 16, 48

Suderinamumo uždavinys 140

– testas 147

Svoris 94

T

t skirstinys 67

Taisyklė:

Čebyšovo 32

sprendimo priėmimo 76, 77, 78, 80, 82, 153, 154, 156

Tankis 33, 112

– chi kvadrato 31 pav.

- eksponentinio dėsnio 115
- gama skirstinio 138
- normaliojo dėsnio 41, 43 (20 pav.), 137
- santykinio dažnio 18, 37
- , savybės 35
- Stjudento 29 pav.

Teorema:

- centrinė ribinė 45, 55, 56, 126
 - charakterizacijos 113
 - empirinių charakteristikų konvergavimo 122
 - empirinių matų konvergavimo 117
 - Glivenko 120, 144
 - Glivenko-Kanteli 120
 - Kolmogorovo 154
 - Kramero 58
 - Pirsono 145
 - Smirnov 156
 - sumos 105
- Testas 141, 142
- suderinamumo (Pirsono) 147
- Tiesinis trendas 88, 90
- Tikimybė 20, 22, 33, 94, 95
- antros rūšies klaidos 147
 - atsitiktinio įvykio 20, 22, 94, 103, 116

- elementariojo įvykio 94
 - , klasikinis apibrėžimas 95
 - pirmos rūšies klaidos 147
- Tikimybės masės funkcija 23
- – –, savybės 25
- Tikimybinių tankio funkcija 35
- Tikimybių skirstinys 108

U**Uždavinys:**

- suderinamumo 140
- skirstinio atitikimo 143, 144

V

Veiksmas be praeities poveikio 112

Vertinimas:

- intervalinis 60
- skirstinių parametrų 115
- taškinis 60

Vidurkis:

- aritmetinis 25
- atsitiktinio dydžio 26
- diskrečiojo dydžio 26
- empirinis 30, 46, 48, 121
- fizikinė interpretacija 27, 39
- imties 25, 29
- kvadratinio nuokrypio 131
- populiacijos 26, 51

Viltis matematinė 26, 53

VARTOJAMŲ SIMBOLIŲ LENTELE

Simbolis	Terminas	Skyrelis, punktas
A	atsitiktinis įvykis	1.6, 2.4, 28.1
$\mathcal{A}$	algebra	26.2
α	reikšmingumo lygmuo	42.7
$1 - \alpha$	pasiklovimo lygmuo	16.1, 45.4
$\mathcal{B}$	Borelio σ -algebra	27.1
ε_i	netiktis (liekana), paklaida	23.2
$\mathbf{E}_\theta$	matematinė viltis pagal skirstinį F_θ	37.2
$\mathcal{F}$	σ -algebra	26.4, 28.1
$F(x)$	skirstinys, skirstinio funkcija	8.8, 30.2
$F_n^*(x)$	empirinis skirstinys	33.1
$f(x)$	tikimybės masės funkcija, tankio funkcija (tankis)	5.4, 8.5
$f(t)$	charakteristinė funkcija	14.5
H_0	nulinė (pagrindinė) hipotezė	19.1, 41.2
H_1	alternatyvioji hipotezė	19.1
θ^*	parametro θ įvertinys	15.7, 35.1, 35.2
$I_X(B)$	atsitiktinio dydžio X indikatorius	32.5
$\mathcal{K}(x)$	Kolmogorovo skirstinys	45.2
$L(\theta)$	logaritminė tikėtinumo funkcija	40.5
$\mu, \mathbf{E}X$	atsitiktinio dydžio X vidurkis, matematinė viltis, populiacijos vidurkis	6.4
$\mu_{\bar{X}}$	įvertinio $\bar{X}$ skirstinio vidurkis	12.6, 12.7, 12.8
M_k^*	empirinis k -tosios eilės momentas	34.1
n	imties dydis, eksperimentų skaičius	2.7, 4.1
N	normalusis atsitiktinis dydis	10.6
$\nu/n, n(A)/n$	santykinis dažnis	1.6, 4.1
$\mathbf{P}$	tikimybinis matas	27.3
$\mathbf{P}(A)$	atsitiktinio įvykio A tikimybė	4.1, 25.2, 25.3, 28.2
$p(x)$	tankis	30.7
$\mathbf{P}_n^*(B)$	empirinis matas	32.4
$\mathcal{R}$	nulinės hipotezės atmetimo (kritinė) sritis	20.5, 20.6
r	koreliacijos koeficientas	22.7, 23.1

Simbolis	Terminas	Skyrelis, punktas
σ	atsitiktinio dydžio standartinis nuokrypis	7.6
$\sigma^2, \mathbf{D}X$	atsitiktinio dydžio X dispersija	7.6
$\sigma^2_{\bar{X}}$	įvertinio $\bar{X}$ skirstinio dispersija	12.6, 12.7, 12.8
s	empirinio standartinio nuokrypio reikšmė	7.4
s^2	imties dispersijos, empirinės dispersijos reikšmė	7.4
S^2	empirinė dispersija (kaip įvertinys)	17.2, 34.1
T	Stjudento statistika	17.2
$t_{\nu, \alpha}$	Stjudento skirstinio su ν laisvės laipsnių lygmens α kritinė reikšmė	17.4
χ^2	atsitiktinis dydis, turintis chi kvadrato skirstinį	18.4, 43.5
$\chi^2_{\nu, \alpha}$	chi kvadrato skirstinio su ν laisvės laipsnių lygmens α kritinė reikšmė	18.4, 43.7
$X, X(\omega)$	atsitiktinis dydis	5.2, 29.2
$x_j, x_j = X(\omega_j)$	atsitiktinio dydžio reikšmė	5.2
$\bar{X}$	empirinis vidurkis (kaip įvertinys)	11.3, 12.1
$\bar{x}$	empirinio vidurkio reikšmė, imties vidurkio reikšmė	6.1, 7.1, 12.1
Z	standartinis normalusis atsitiktinis dydis	10.5
ω_j	elementarusis įvykis, populiacijos elementas	2.1, 2.7, 25.1
Ω	elementariųjų įvykių erdvė, populiacija	2.1, 2.7, 25.1, 28.1
Ω	imčių erdvė	2.7
Ω	būtinasis įvykis	2.5
$(\Omega, \mathbf{P})$	diskretusis tikimybinis modelis	25.3
$(\Omega, \mathcal{F})$	mačioji erdvė	27.2, 28.1
$(\Omega, \mathcal{F}, \mathbf{P})$	tikimybinė erdvė	28.2
$\emptyset$	negalimasis įvykis	2.6, 28.1
$\Rightarrow$	silpnas konvergavimas	36.2