Liet. matem. rink., T. 40, spec. nr., 111–116

Pseudo-atsitiktiniai skaičiai statistinėse programinėse sistemose

Vitalija RUDZKIENĖ (VGTU, LTA)
el. paštas: vital@fm.vtu.lt

1. Įvadas

Atsitiktiniai skaičiai naudojami skaitiniuose metoduose, gamtinių ir sociologinių reiškinių modeliavime, sprendimo priėmimo metoduose, kompiuteriniuose žaidimuose ir kt. „Tikri" atsitiktiniai skaičiai gaunami fiziniais signalų generatoriais, kurie naudoja natūralius atsitiktinių triukšmų šaltinius (elektroninių prietaisų triukšmus, radioaktyvų skilimą ir t.t.). Programiškai kompiuterinėmis programomis generuojami atsitiktiniai skaičiai vadinami pseudo-atsitiktiniais skaičiais. Praktika rodo, kad nuo atsitiktinių skaičių kokybės tiesiogiai priklauso gaunami rezultatai. Todėl pseudo-atsitiktinių skaičių generatoriai yra testuojami, tačiau nustatyti, ar seka atsitiktinė nėra paprasta. Praktikoje atsitiktinių dydžių generavimui dažniausiai yra naudojamos statistinės programinės sistemos (statistiniai paketai). Šiuo metu pagal Tarptautinio statistikos instituto duomenis, statistinių paketų skaičius artėja prie 1000. Lietuvoje yra populiarūs statistiniai paketai Statistica, Statgraphics, SPSS, SAS. Visi šie paketai turi neparametrinį modulį su pseudo-atsitiktinių skaičių generavimo procedūromis. Tačiau dažnai testuojant pseudo-atsitiktinių skaičių seką, sugeneruotą pagal pasirinktą pasiskirstymo dėsnį, pavyzdžiui, eksponentinį, pasirodo, kad šios sekos savybės neatitinka nurodyto pasiskirstymo dėsnio. Šiame darbe analizuojamos ir lyginamos pseudo-atsitiktinių skaičių sekos, gautos statistinių paketų pagalba.

2. Pseudo-atsitiktinių skaičių generatoriai

Iš esmės kompiuterinių programų skaičiavimo rezultatai yra determinuoti ir naują rezultatą galime gauti tik pakeitus programos duomenis. Vis dėlto yra imanoma, kad kompiuteris atspausdintų skaičių seką $\{u(i)\}$, kuri pagal standartinius statistinius testus atrodytų esanti nepriklausomų atsitiktinių dydžių seka. Algoritmai, kurie sukuria tokius skaičius, vadinami pseudo-atsitiktinių skaičių generatoriais. Paprastai tokie generatoriai yra rekurentiniai:

$$u_i = d(u_{i-1}, u_{i-2}, ..., u_{i-p}). \tag{1}$$

Nurodžius pradines reikšmes $u_0, u_{-1}, ..., u_{-p+1}$, būsimas reikšmės apibrėžia funkcija d. Kadangi kompiuterio skaičiuojamų reikšmių aibė yra baigtinė, pseudo-atsitiktinių skaičių

generatoriai sukuria ciklišKai pasikartojančias reikšmes. Pasirinkdamas funkciją d, vartotojas tuo pačiu pasirenka ir imčių erdvę, kurios kiekvienas elementas yra ciklo seka. Pagrindinis uždavinys yra rasti tokią funkciją d, kuriai atitiktų ir imčių erdvė su reikiamomis savybėmis, ir kad pradinių reikšmių $u_0, u_{-1}, ..., u_{-p+1}$ parinkimas būtų lengvas ir paprastas. Labiausiai paplitęs yra Lehmer tiesinis kongruentinis generatorius, kurio pavidalas

$$u_i = (au_{i-1} + c)\mathrm{mod}(m), \tag{2}$$

ir multiplikatyvinis kongruentinis generatorius, kuris yra atskiras tiesinio kongruentinio generatoriaus atvejis, kai $c = 0$. Šalia privalumų, šie generatoriai turi eilę trūkūmų. Pagrindinis trūkumas yra tas, kad sugeneruotas k-matis vektorius

$$u_{i+1}, u_{i+2}, \ldots, u_{i+k} \tag{3}$$

gali padengti tik mk tinklo taškų k -matėje erdvėje. O blogiausia, kad padengti taškai guli ne daugiau kaip $(k!m)1/k$ hiperplokštumose. Kuo k didesnis ir m mažesnis, tuo ribos siauresnės [1]. Pastaraisiais metais buvo pasiūlyta keletas naujo tipo atsitiktinių skaičių generatorių, tokių, kaip inversinis kongruentinis generatorius

$$u_i = (a(u_{i-1})^{-1} + c)\mathrm{mod}\,(m) \tag{4}$$

ir Tausworthe atgalinio ryšio-poslinkio-kaupimo generatorius apibrėžiamas lygtimi

$$u_i = (c_1u_{i-1} + c_2u_{i-2} + +c_pu_{i-p})\mathrm{mod}(2), \tag{5}$$

kur $c_1, c_2, ..., c_p$ yra dvejetainiai ir $c_p = 1$.

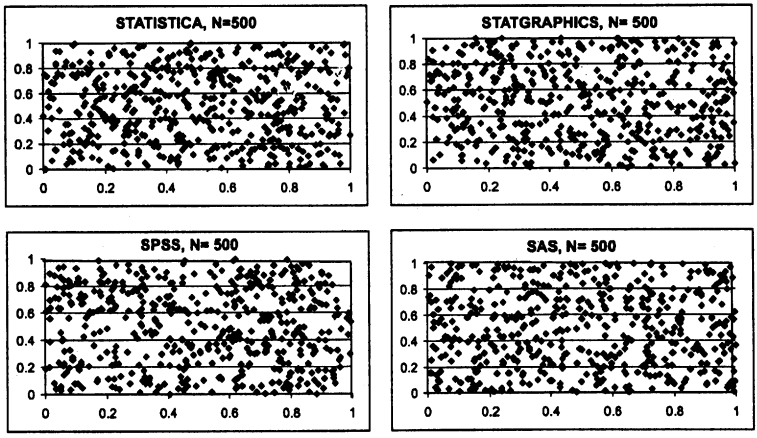
Generatorių pasirinkimą dažniausiai apsprendžia programinės įrangos kūrėjai. Net jei vartotojas ir gali pasirinkti, tai šis pasirinkimas galimas tik iš sąlyginai mažos generatorių aibės. Programinės įrangos dokumentacijoje paprastai nenurodoma, koks generatoriaus algoritmas yra naudojamas.

3. Pseudo-atsitiktinių skaičių sekų, sugeneruotų statistiniais paketais Statistica, Statgraphics, SPSS, SAS savybių tyrimas

Kaip žinoma, pseudo-atsitiktiniai dydžiai generuojami dviem etapais: 1) gaunami tolygiai vienetiniame intervale pasiskirstę atsitiktiniai dydžiai; 2) šie atsitiktiniai dydžiai transformuojami į atsitiktinius dydžius su pageidaujama pasiskirstymo funkcija. Kadangi visi tikimybiniai skirstiniai yra gaunami iš tolygaus intervale $[0, 1]$, tai šiame darbe apsiribosime tolygiai pasiskirsčiusių vienetinime intervale sekų tyrimu. Reikia pripažinti, kad šiuo metu statistinė hipotezių apie procesų atsitiktinumą tikrinimo teorija nėra pakankamai gerai išvystyta, ypač alternatyvų atžvilgiu. Nustatyti, ar seka atsitiktinė, yra gana sudėtinga. Paprastai šiam tikslui yra naudojami χ^2 arba Kolmogorovo-Smirnovo kriterijai, kurie leidžia nustatyti ar stebimi dažniai atitinka tiriamą pasiskirtymo dėsnį (nulinė

hipotezė) ir serijų kriterijus, tikrinintis nulinę hipotezę apie atsitiktinį sekos pobūdį. Šiuos kriterijus ir panaudosime statistiniais paketais sugeneruotų sekų analizei.

Tyrimui buvo panaudota po dešimt pseudo-atsitiktinių skaičių sekų ilgio $N = 100, 500, 1000$ ir 5000, gautų statistiniais paketais Statistica'98, Statgraphics 2.1, SPSS 6.1.3, SAS 6.12. Nupiešus gautų dydžių porų vaizdus, pastebimas jų jungimasis į kreives. Kadangi gauti vaizdai buvo panašūs visų ilgių sugeneruotoms sekoms, parodysime tik dvimačius vaizdus pseudo-atsitiktinių dydžių (u_i, u_{i-1}), $i = 2, 4, 6, \ldots$ gautų skirtingais paketais esant sekos ilgiui $N = 500$.



1 pav. Dvimačiai pseudo-atsitiktiniai dydžiai (u_i, u_{i-1}), $i = 2, 4, 6, \ldots$, gauti skirtingais statistiniais paketais.

1 lentelė

χ^2 kriterijaus reikšmingumo lygmenų aritmetiniai vidurkiai įvairiems sekų ilgiams N.

Imties dydis N	Statistica	Statgraphics	SPSS	SAS
100	0.3857	0.34153	0.44	0.4423
500	0.4762	0.6801	0.37043	0.4155
1000	0.3719	0.7512	0.4475	0.52645
5000	0.6742	0.9892	0.40834	0.5537

Toliau tikrinama nulinė hipotezė H_0 apie sugeneruotų sekų tolygų pasiskirstymą intervale $[0, 1]$. Visi keturi čia tiriami paketai turi po du suderinamumo kriterijus: χ^2 ir Kolmogorovo-Smirnovo, kuriuo ieškomas didžiausias absoliutus skirtumas tarp stebimos ir teorinės pasiskirstymo funkcijos. Mes naudosimės χ^2 kriterijumi. Kiekvienai sekai buvo suskaičiuotos χ^2 kriterijaus reišmės ir jas atitinkantis reikšmingumo lygmuo p. Skaičiavimas buvo atliktas paketu Statistica'98. Grupavimo intervalų skaičius buvo paliktas toks, kokį parinkdavo šis paketas. Kai $N = 100$, buvo grupuota į 10 intervalų, kai $N = 500$, 1000 ir 5000 grupuota į 22 intervalus. Po to iš dešimties kiekvieno ilgio sekų buvo surasti χ^2 kriterijaus reikšmingumo lygmens p aritmetiniai vidurkiai (1 lentelė).

Akivaizdu, kad hipotezės H_0 atmesti negalime. Kai sekų ilgiai $N = 100, 500$ ar 1000 kriterijaus χ^2 įverčiai yra panašaus dydžio. Išsisikiria tik Statgraphics 2.1 paketu gauta vidutinė reikšmė 0.9892 kai $N = 5000$. Kadangi šios sekos dar bus tiriamos toliau, šios reikšmės atsiradimo priežastis galėtų paaiškėti vėliau.

Kad nustatyti, ar reikšmingi skirtumai tarp skirtingais statistiniais paketais generuojamų pseudo-atsitiktinius dydžių, turime patikrinti nulinę hipotezę, kad šias sekas atitinkančių kriterijaus χ^2 reikšmių aritmetiniai vidurkiai yra vienodi, t.y.:

$$\mu_{Statistica} = \mu_{Statgraphics} = \mu_{SPSS} = \mu_{SAS}. \tag{6}$$

Tyrimui panaudosime dispersinės analizės metodą. Tiksliausi dispersinės analizės rezultatai gaunami, kai tiriamos imtys yra normaliai pasiskirsčiusi ir turi vienodas dispersijas. Mūsų atveju žinome, kad imtis pasiskirsčiusi pagal χ^2 dėsnį, o dispersijos yra vienodos, visais atvejais sumuojant buvo laikomasi vienodo intervalo skaičiaus, tuo pačiu ir df – laisvės laipsnių skaičius yra toks pat, o χ^2 skirstinio dispersija yra $2df$. Žinoma, kad dispersinės analizės kriterijus yra gana robastinis imties pasiskirstymo dėsnio atžvilgiu ir nukrypimai nuo normališkumo neturi didelės įtakos rezultatų tikslumui [2]. Skaičiavimus atliksime statistinio paketo Statistica'98 moduliu ANOVA. Pradžioje tirsime hipotezę kad, esant vienodam sekos ilgiui, visų sekų kriterijaus χ^2 reikšmės turi tą patį vidurkį. Dispersinės analizės rezultatai pateikti 2 lentelėje.

2 lentelė

Pseudo-atsitiktinių dydžių sekų, sugeneruotų statistiniais paketais Statistica'98, Statgraphics 2.1, SPSS 6.1.3, SAS 6.12 dispersinės analizės rezultatai.

Sekos ilgis N	F kriterijaus reikšmė	Reikšmingumo lygmuo p
100	0.63	0.60
500	2.67	0.06
1000	4.1	0.013
5000	9.85	0.00007

3 lentelė

Pseudo-atsitiktinių dydžių sekų, sugeneruotų statistiniais paketais Statistica'98, SPSS 6.1.3, SAS 6.12 dispersinės analizės rezultatai (be Statgraphics 2.1).

Sekos ilgis N	F kriterijaus reikšmė	Reikšmingumo lygmuo p
100	0.22	0.802
500	0.92	0.408
1000	0.72	0.49
5000	0.58	0.566

Kaip matome iš dispersinės analizės rezultatų, esant sekos ilgis $N = 1000$ ir reikšmingumo lygmeniui $p = 0.05$ negalime teigti, kad visų sekų kriterijaus χ^2 reikšmių aritmetiniai vidurkiai yra vienodi. Kai sekos ilgis $N = 5000$, tikimybė, kad hipotezė teisinga sumažėja iki 0.00007. Matome, kad iš visų sugeneruotų sekų išsiskiria sekos, generuotos Statgraphics paketu. Patikrinkime, ar pasikeis dispersinės analizės rezultatai, atmetus sekas, generuotas Statgraphics paketu (3 lentelė).

Iš rezultatų, pateiktų 3 lentelėje matome, kad sekas, sugeneruotas statistiniais paketais Statistica'98, SPSS 6.1.3, SAS 6.12, galima laikyti turinčias vienodus kriterijaus χ^2 reikšmių aritmetinius vidurkius. Taigi, iš visų sekų išsiskiria tik statistiniu paketu Statgraphics 2.1 generuotos sekos ir tik kai sekos ilgis didesnis už 500.

Jeigu manome, kad rezultatai galėjo būti iškraipyti todėl, kad tiriamos sekos pasiskirsčiusios pagal χ^2, o ne pagal normalųjį dėsnį, gautus rezultatus galima patikrinti ranginiu Kruskal-Wallis kriterijumi (paketas SPSS). šiuo kriterijumi gauti rezultatai mažai skiriasi nuo dispersinės analizės rezultatų (4, 5 lentelės).

4 lentelė

Pseudo-atsitiktinių dydžių sekų, sugeneruotų statistiniais paketais Statistica'98, Statgraphics 2.1, SPSS 6.1.3, SAS 6.12 Kruskal-Wallis kriterijaus rezultatai.

Sekos ilgis N	Kruskal-Wallis kriterijaus reikšmė	Reikšmingumo lygmuo p
100	0.64	0.89
500	7.11	0.068
1000	10.87	0.0124
5000	22.61	0.0000

5 lentelė

Wald-Wolfowitz runs kriterijaus reikšmingumo lygmenų aritmetiniai vidurkiai įvairiems sekų ilgiams N.

Imties dydis N	Statistica	Statgraphics	SPSS	SAS
100	0.627	0.550	4.844	0.474
500	0.497	0.605	0.564	0.491
1000	0.497	0.552	0.373	0.682
5000	0.506	0.545	0.506	0.391

Generuotų sekų atstiktinumo tyrimui buvo panaudotas plačiausiai paplitęs Wald-Wolfowitz serijų (runs) kriterijus. Šis kriterijus realizuotas daugelyje statistinių paketų. 5 lentelėje pateikti Wald-Wolfowitz serijų kriterijaus aritmetiniai vidurkiai. Mažiausias vidutinis patikimumo lygmuo p yra 0.373, gautas tyriant sekas, sugeneruotas SPSS paketu. Iš gautų kriterijaus reikšmių ir gana didelių reikšmingumo lygmenių p matome, kad hipotezės apie sekų atsitiktinumą atmesti negalime.

Wald-Wolfowitz runs kriterijaus reikšmės pasiskirsčiusios pagal normalųjį dėsnį, taigi, tenkinamos visos prielaidos dispersinei analizei atlikti. Visais atvejais gauti dispersinės analizės rezultatai su gana dideliu reikšmingumo lygmeniu p (bendras vidutinis reikšmingumo lygmuo $p = 0.52$) neleidžia atmesti nulinės hipotezės, kad visos sekos turi panašius Wald-Wolfowitz runs kriterijaus reikšmingumo lygmenų aritmetinius vidurkius.

4. Išvados

1. Pseudo-atsitiktinių skaičių sekų, gautų statistiniais paketais Statistica'98, Statgraphics 2.1, SPSS 6.1.3, SAS 6.12 skirstinio savybės buvo tiriamos χ^2 kriterijumi. Gauti rezultatai rodo, kad tolygaus intervale $[0, 1]$ skirstinio savybes geriausiai atitinka sekos, generuotos Statgraphics 2.1 statistiniu paketu ir tik esant pakankamai ilgai sekai, kai jos ilgis $N \geqslant 500$.
2. Iš dispersinės analizės rezultatų galime teigti, kad pseudo-atsitiktinių skaičių sekų, gautų statistiniais paketais Statistica'98, SPSS 6.1.3, SAS 6.12, tolygaus pasiskirtymo dėsnio intervale $[0, 1]$ savybės yra panašios ir mažai keičiasi, ilgėjant sekoms.
3. Generuotų sekų atstiktinumo tyrimui buvo panaudotas Wald-Wolfowitz serijų kriterijus. Dispersinės analizės rezultatai su gana dideliu reikšmingumo lygmeniu p (bendras vidutinis reikšmingumo lygmuo $p = 0.52$) neleidžia atmesti nulinės hipotezės, kad visos sekos turi panašius Wald-Wolfowitz runs kriterijaus reikšmingumo lygmenų aritmetinius vidurkius.

Literatūra

[1] D.P. Heyman and M.J. Sobel, *Stochastic Models*, North-Holland (1990).
[2] H. Kohler, *Essentials of Statistics*, Scott, Foresman and Company, Illinois (1988).

Quasi-random numbers in some statistical systems

V. Rudzkienė

There are many ways to get random numbers. Some methods include making hardware devices that generate noise, observing cosmic ray flux and etc. Pseudo-random numbers come from mathematical functions and algorithms that provides such numbers called pseudo-random generators. The generators chosen is often that provided by the software developer. The most popular way for getting pseudo-random numbers is by statistical programming systems. In Lithuania these statistical systems are in use: Statistica, Statgraphics, SAS, SPSS. Statistical properties of random sequences, generated by these statistical systems, are investigated in this paper.