Suderinamumo su normaliuoju skirstiniu hipotezių tikrinimo kriterijų kūrimas ir galingumo lyginamasis tyrimas naudojant Monte Karlo metodą

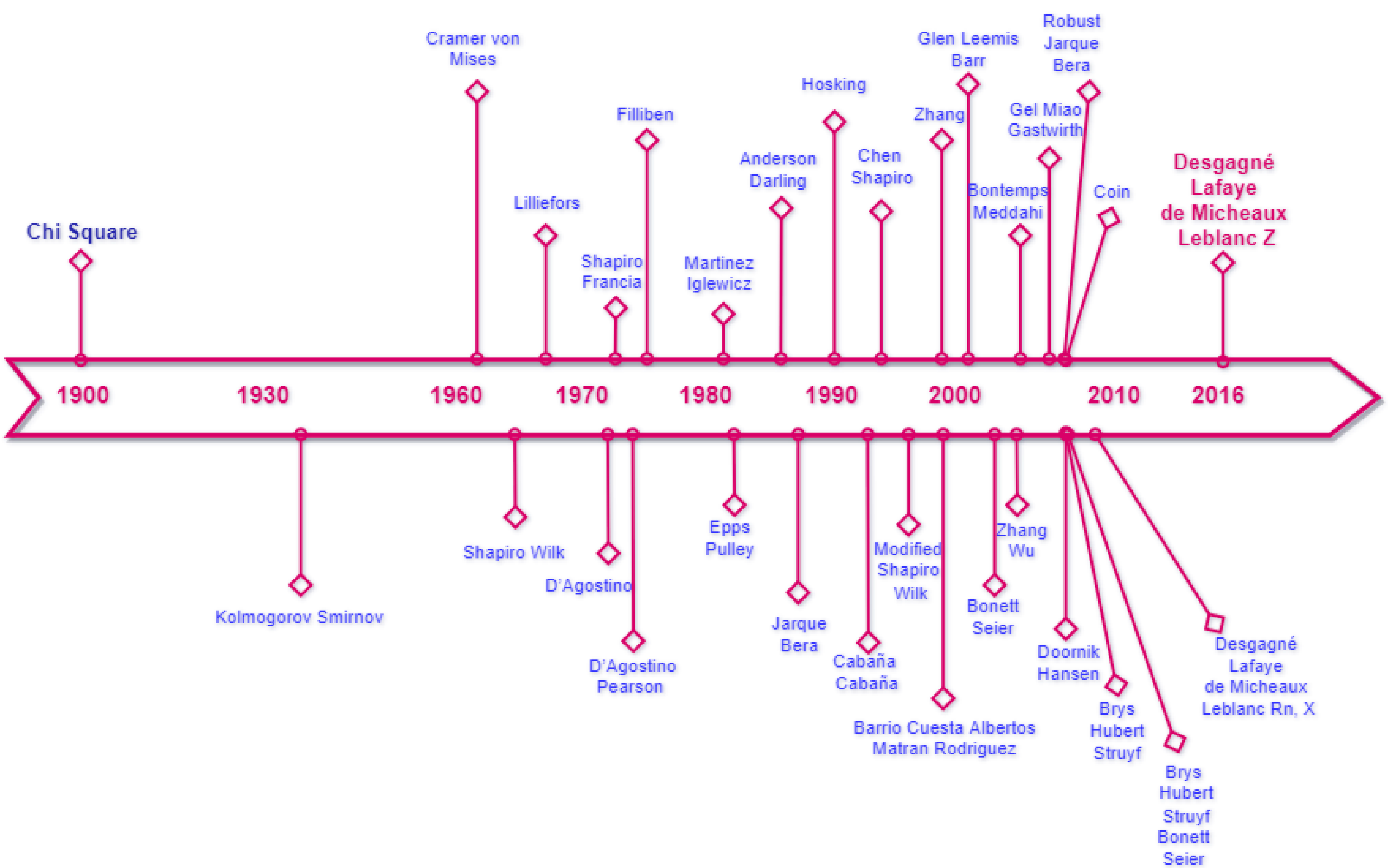
2022 m. vasario 15 d.

Doktorantė: Jurgita Arnastauskaitė

Vadovas: doc. dr. Tomas Ruzgas

Problematika

Duomenų tyrybos problema, su kuria susiduriama daugelyje tyrimų sričių, yra poreikis įvertinti, imties pasiskirstymą.



Tikslas ir uždaviniai

Sukurti ir ištirti suderinamumo hipotezės tikrinimo vienmačius ir daugiamačius kriterijus, kurie būtų efektyvūs normalumo prielaidos atveju.

1. Ištirti pagrindines problemas, sprendimus bei kliūtis atliekant suderinamumo hipotezių tikrinimą.
2. Sudaryti naujus vienmatį ir daugiamatį suderinamumo hipotezės tikrinimo kriterijus, skirtus normalumo prielaidos patikrinimui.
3. Atlikti sudarytų ir populiarių vienmačių ir daugiamačių suderinamumo hipotezės tikrinimo kriterijų galios lyginamąją analizę, moksliniuose tyrimuose taikomų duomenų pasiskirstymų atvejais.
4. Pritaikyti sudarytus suderinamumo hipotezės tikrinimo kriterijus realių duomenų rinkiniams.

Naujumas

Siekiant išspręsti normalumo prielaidos tikrinimo problemą, buvo atlikti išsamūs vienmačių ir daugiamačių kriterijų palyginimo tyrimai įvairiems duomenų rinkiniams. Pasiūlyti vienmatis ir daugiamatis kriterijai, kurie yra galingi kitų kriterijų konkurentai.

Tyrimo objektas, subjektas ir matas

- Tyrimo **objektas** yra suderinamumo hipotezė.
- Tyrimo **subjektas** yra suderinamumo hipotezės tikrinimo kriterijus.
- Tyrimo **matas** yra suderinamumo hipotezės tikrinimo kriterijaus galia.

Kriterijaus galia

Nulinės hipotezės atmetimo tikimybė

$$\boldsymbol{\beta}^*(\theta) = E_{\theta\varphi}(X), \theta \epsilon \Xi,$$

vadinama kriterijaus galia.

čia $\varphi = \begin{cases} 0, & x \in K, \\ 1, & x \in A \end{cases}$ yra kriterijaus funkcija,

$K -$ kritinė sritis, $A -$ priėmimo sritis.

Ginamieji teiginiai

- Pasiūlytas N-metrikos teorija paremtas, suderinamumo hipotezės tikrinimo kriterijus, kuris yra galingesnis už kitus galingus vienmačius kriterijus dideliems imties dydžiams.
- Pasiūlytas pasiskirstymo tankių skirtumo vertinimu ir apvertimo formulės taikymu paremtas, daugiamatis suderinamumo hipotezės tikrinimo kriterijus, kuris pasižymi ženkliai didesne galia nei kiti galingiausi kriterijai simetrinių ir mišriųjų skirstinių grupėms.

VIENMATIS ATVEJIS

Vienmačiai kriterijai

Pasiūlytas kriterijus (N1)

Esant nulinės hipotezės statistikai $T_n = -n\int_0^1\int_0^1 K(x)d(F_n^*(x)-x)$ asimptotinis pasiskirstymas bus toks pats kaip kvadratinės formos:

$$T_n = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\frac{a_{kj}}{\pi^2 kj}\xi_k\,\xi_j,$$

čia ξ_k – nepriklausomas atsitiktinis dydis pasiskirstęs pagal standartinį normalųjį skirstinį

$$a_{kj} = -2\int_0^1\int_0^1 K(x)\,d\sin(\pi kx)\,,$$

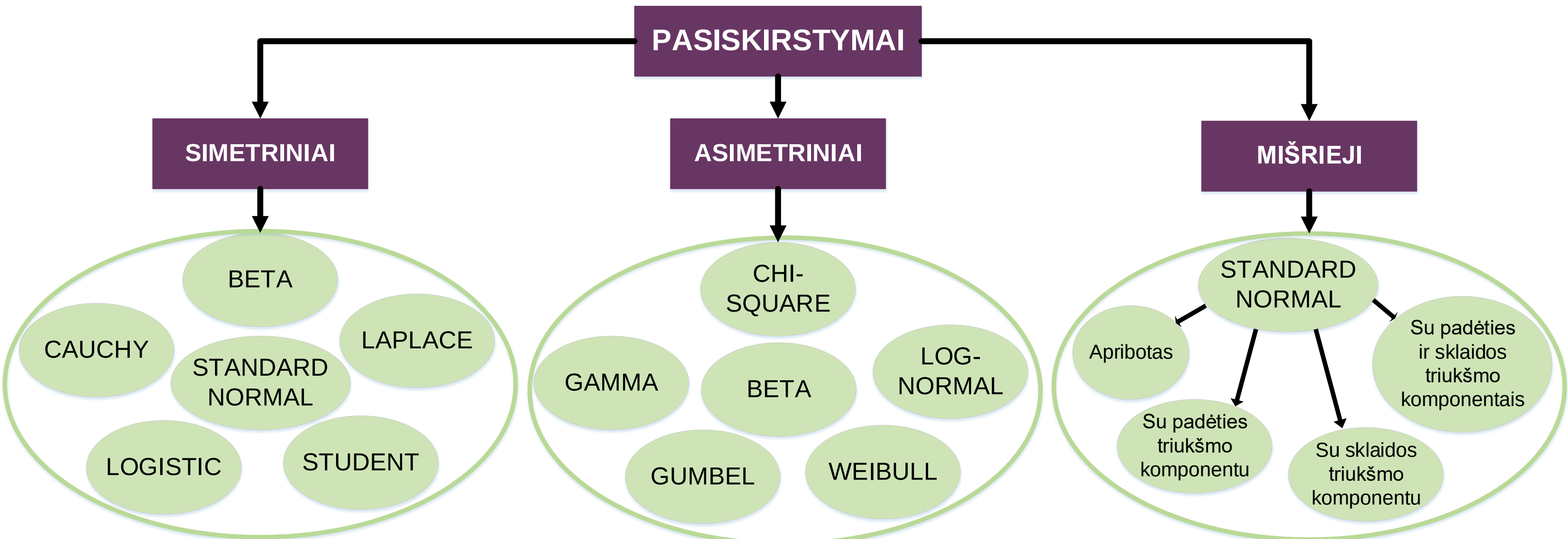
čia $K(x)=\varphi\big(\bar{g}(x)\big)\bar{g}'(x)$ – pasiūlyta branduolio funkcija,

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}},$$

$$\bar{g}(x)=x\big(c+|x|^b\big)^a,(a>0,b\geq 1,c>0),$$

$$\bar{g}'(x)=\big(c+|x|^b\big)^a+a|x|\big(c+|x|^b\big)^{a-1}b|x|^{b-1}.$$

Vienmačiai pasiskirstymai



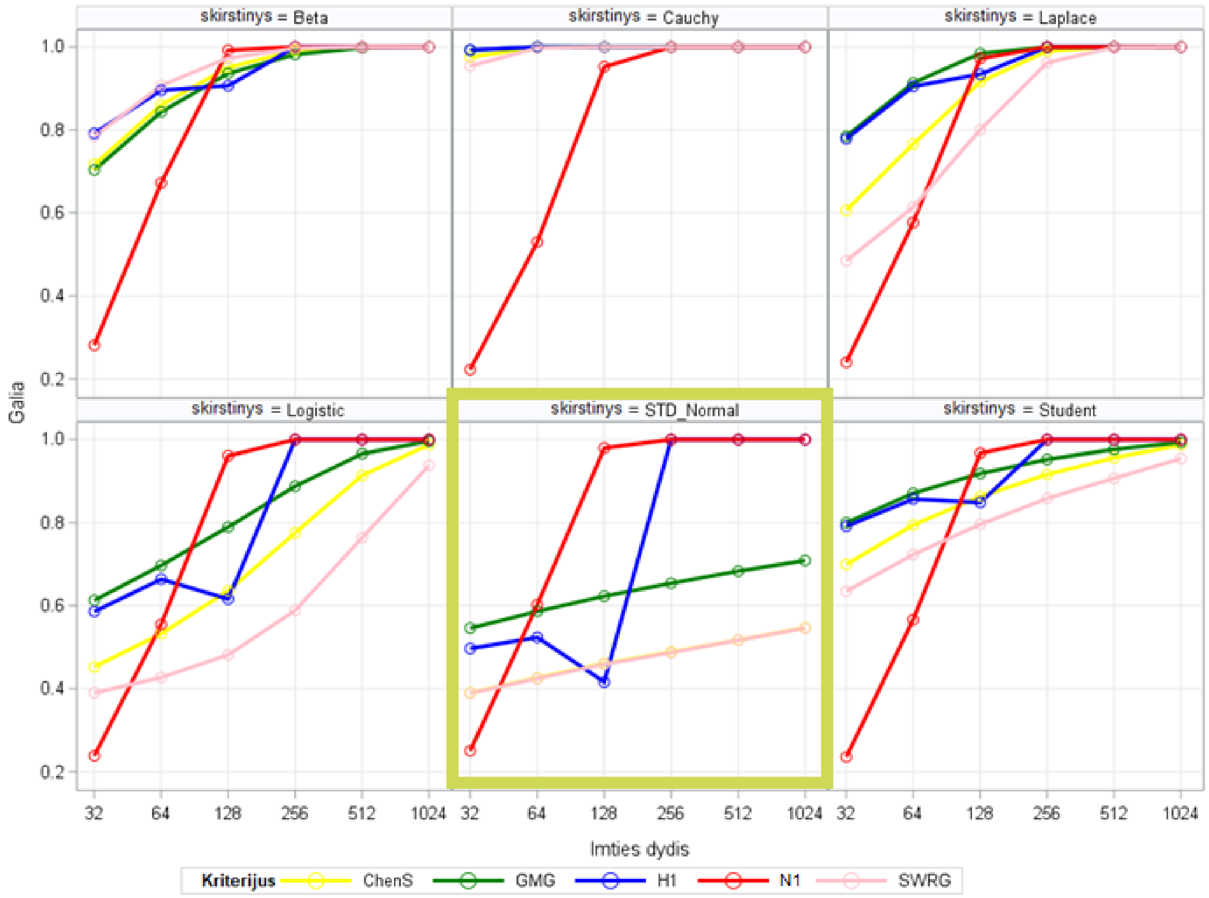
Modeliavimo tyrimas

Tyrimas atliktas taikant paskirstytuosius skaičiavimus su:

- **3** skirstinių grupėmis;
- **6** imčių tūriais ($32, 64, 128, 256, 512, 1024$);
- **42** kriterijų (įskaitant ir naujai sukonstruotą);
- **1 000 000** nepriklausomų imčių.

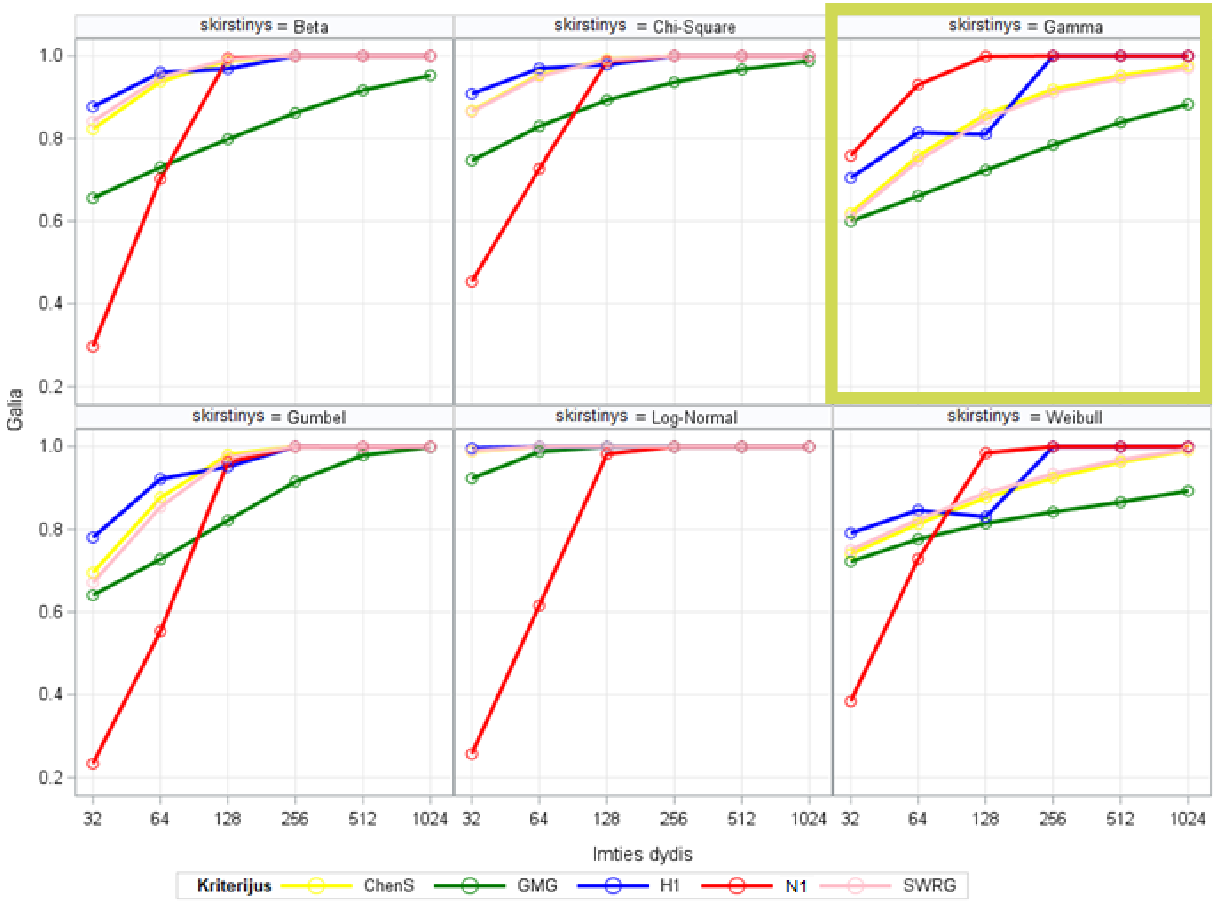
Modeliavimo tyrimas

Simetrinių skirstinių grupė



Modeliavimo tyrimas

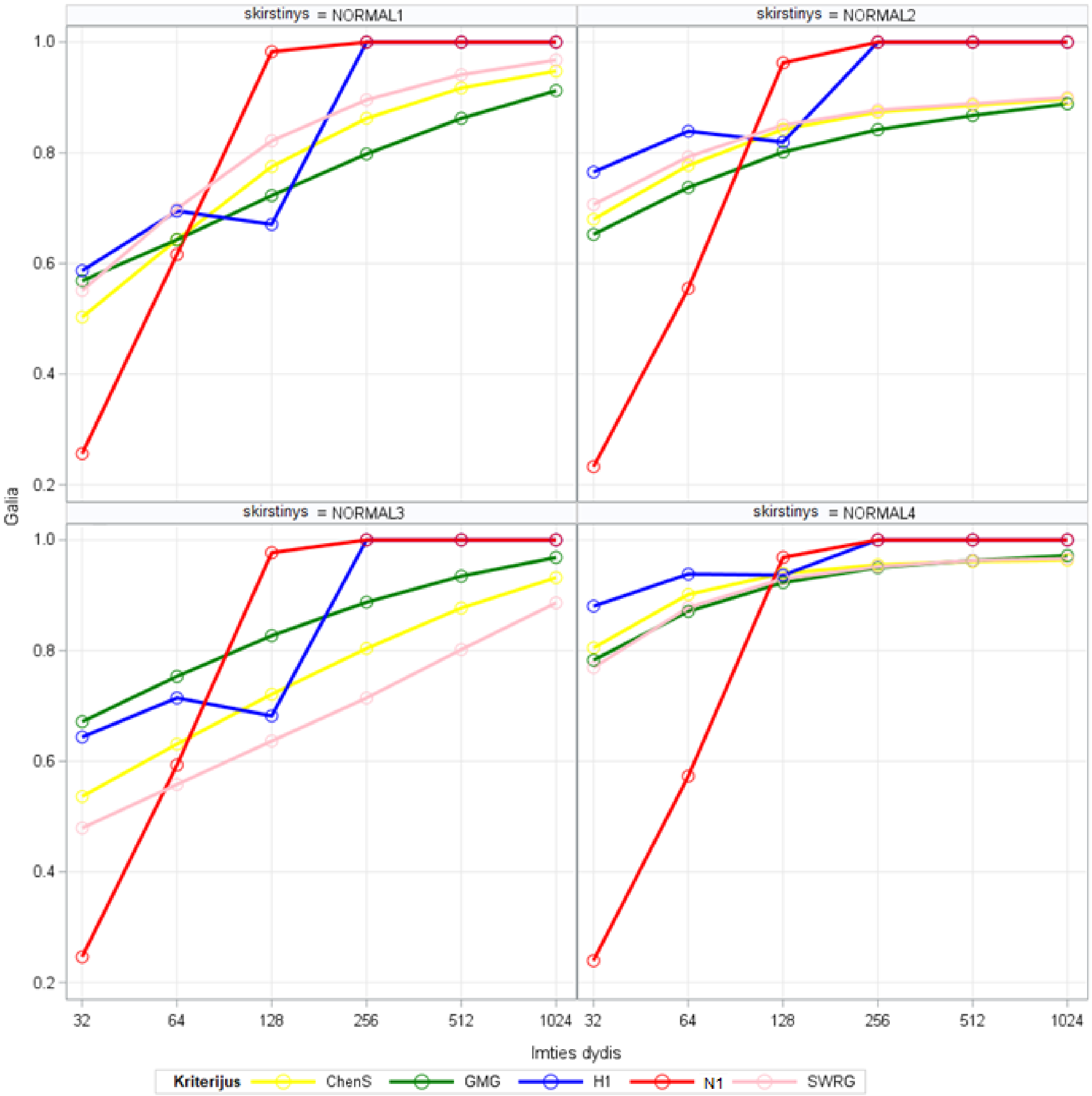
Asimetrinių skirstinių grupė



ktu
1922

Modeliavimo tyrimas

Modifikuotų normaliųjų skirstinių grupė



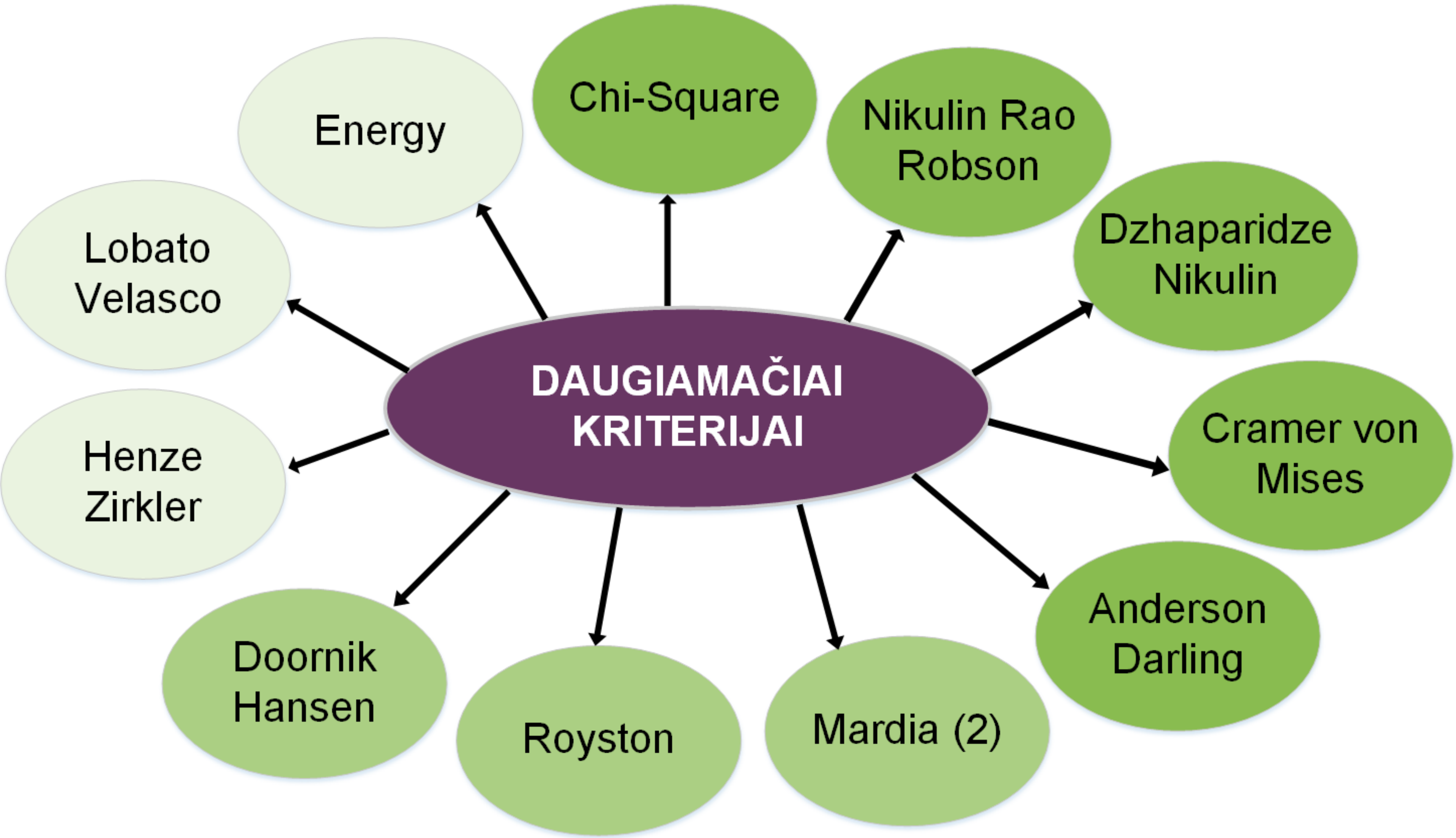
Mažiausias imties dydis

Buvo atliktas papildomas tyrimas, siekiant įvertinti mažiausią imties dydį, kai *N1* kriterijus tampa galingesnis simetriškų, asimetrinių ir modifikuotų normalių skirstinių grupėms. *Hosking1* ir *N1* kriterijai buvo taikomi 80, 90, 100, 105, 110, 115 dydžių duomenų rinkiniams.

Nr.	Skirtinys	Skirstinio grupės	Minimalus imties dydis (n)
1.	Standartinis normalusis	Simetrinė	46
2.	Beta	Simetrinė	88
3.	Koši	Simetrinė	257
4.	Laplaco	Simetrinė	117
5.	Logistinis	Simetrinė	71
6.	Studento	Simetrinė	96
7.	Beta	Asimetrinė	108
8.	Chi-kvadrato	Asimetrinė	123
9.	Gama	Asimetrinė	**< 32**
10.	Gumbelio	Asimetrinė	125
11.	Log-normalusis	Asimetrinė	255
12.	Veibulo	Asimetrinė	65
13.	Normal1	Modifikuota	70
14.	Normal2	Modifikuota	93
15.	Normal3	Modifikuota	72
16.	Normal4	Modifikuota	117

DAUGIAMATIS ATVEJIS

Daugiamačiai kriterijai



Pasiūlytas kriterijus (N2)

Pasiūlytas normalumo hipotezės kriterijus yra paremtas integralu:

$$\mathcal{T} = \int_{-\infty}^{\infty} \left| f(t) - \hat{f}(t) \right| dG(z), \qquad (1)$$

čia z – standartizuoti dydžiai, $\hat{f}(z)$ įvertinys

$$\hat{f}(x) = \frac{A(d)}{\#T} \sum_{\tau \in T} \left[\sum_{k=1}^{\hat{q}_\tau} \hat{p}_{k,\tau} I_{d-1} \left(\frac{\widehat{m}_{k,\tau} - \tau' x}{\sqrt{\hat{\sigma}^2_{k,\tau} + 2h}} \right) \left(\hat{\sigma}^2_{k,\tau} + 2h \right)^{-\frac{d}{2}} + \frac{2\hat{p}_{0,\tau}}{b-a} J_{d-2} \left(\frac{a + b - 2\tau' x}{2\sqrt{2h}}, \frac{b-a}{2\sqrt{2h}} \right) \cdot (2h)^{-\frac{d-1}{2}} \right],$$

čia # žymi aibės T elementų skaičių,

$$A(d) = \frac{\left(V_d(1)\right)'_R}{(2\pi)^d} = \frac{d 2^{-d} \pi^{-\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)},$$

$$I_j(y) = \mathrm{Re}\left[\int_0^\infty e^{iyz - z^2/2} z^j dz\right],$$

kai

$$J_j(y,z) = Re\left[\int_0^\infty e^{iyu - u^2/2} \cdot \sin zu \cdot u^j \, du\right].$$

Pasiūlytas kriterijus (N2)

ktu
1922

$G(z)$ pasirinkimui įtakos turi trys savybės:

- $G(t)$ priskiria $\left|f(z) - \hat{f}(z)\right|$ didelį svorį, kai pasiskirstymo tankis susijęs su alternatyvia hipoteze.
- $G(z)$ suteikia didelį svorį, kai $\hat{f}(z)$ statistika yra palyginti tikslus $f(z)$ įvertinys.
- $G(z)$ yra toks, kad integralas (1) yra uždaros formos.

Neparametriniams metodams pirmos dvi savybės tenkinamos tinkamai parinkus tankio įvertinį $\hat{f}(z)$, be to tai suteikia uždarą (1) integralo formą:

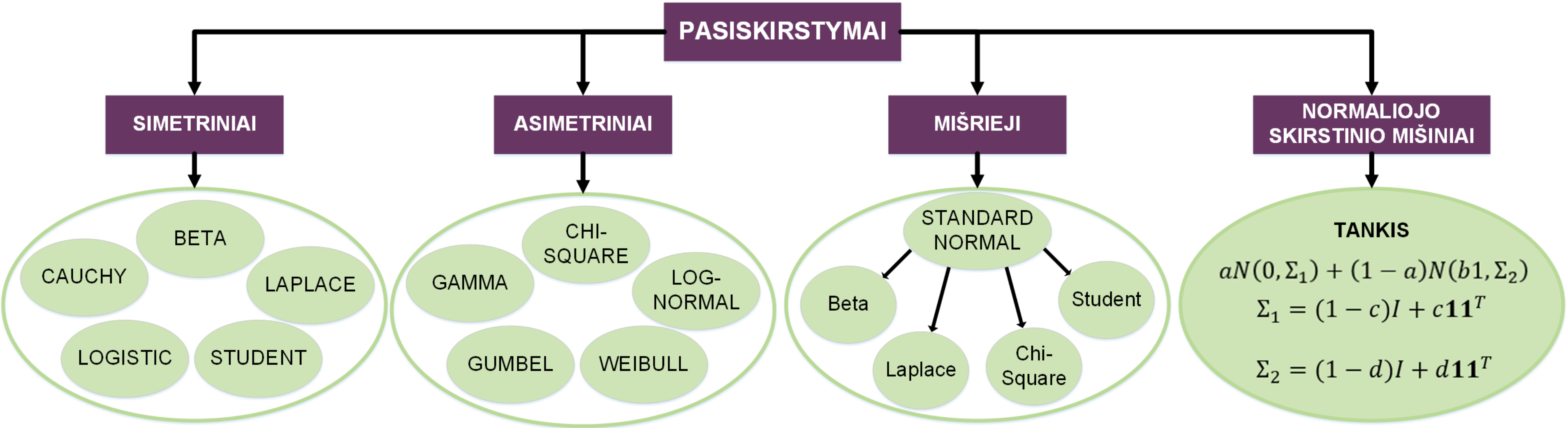
$$\mathcal{T} = n^{-1} \sum_{t=1}^{n} \left|f(z) - \hat{f}(z)\right|,$$

$\mathcal{T}$ nepriklauso nuo saikingo imties tūrio (≥32), tačiau priklauso nuo duomenų dimensijos. Pasirinkta naudoti kriterijaus statistiką $\mathcal{T}^* = -\log\{\mathcal{T}\}$, kuri pasižymėjo mažiausiu jautrumu nuo žvalgomojo tyrimo taškų. Statistiką $\mathcal{T}^*$ galima aproksimuoti Johnsono SU skirstiniu, kurio tankio funkcija apibrėžiama:

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}} g'\left(\frac{x-\xi}{\lambda}\right) \exp\left(-0{,}5\left[\gamma + \delta g\left(\frac{x-\xi}{\lambda}\right)\right]^2\right),$$

čia $g(y) = \ln\left[y + \sqrt{y^2+1}\right]$, kai $x\epsilon(-\infty, +\infty)$. Parametrai: $\delta -$ formos, $(\delta > 0)$, $\gamma -$ formos, $\xi -$ padėties, $\lambda -$ sklaidos, $(\lambda > 0)$.

Daugiamačiai pasiskirstymai

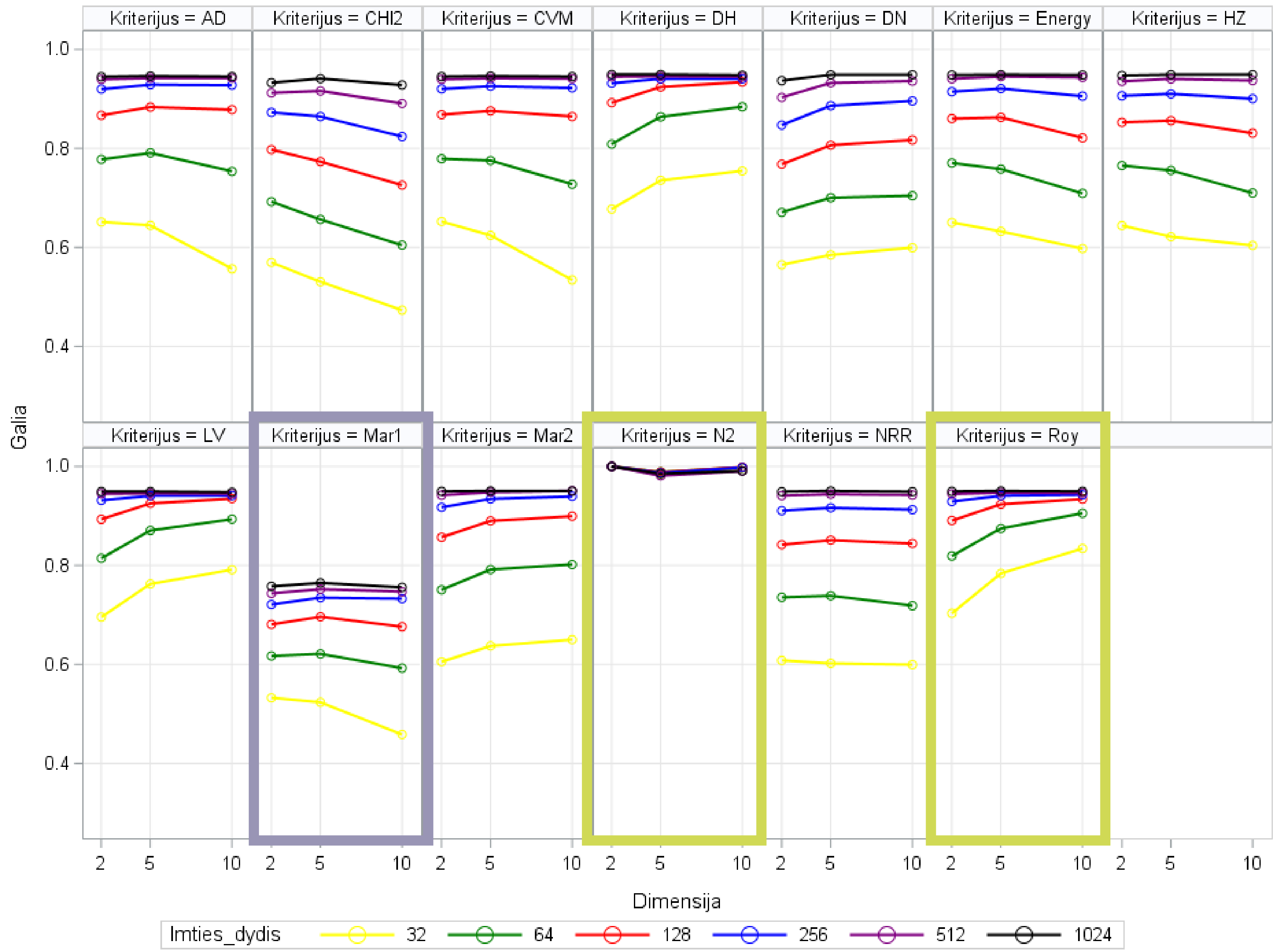


Modeliavimo tyrimas

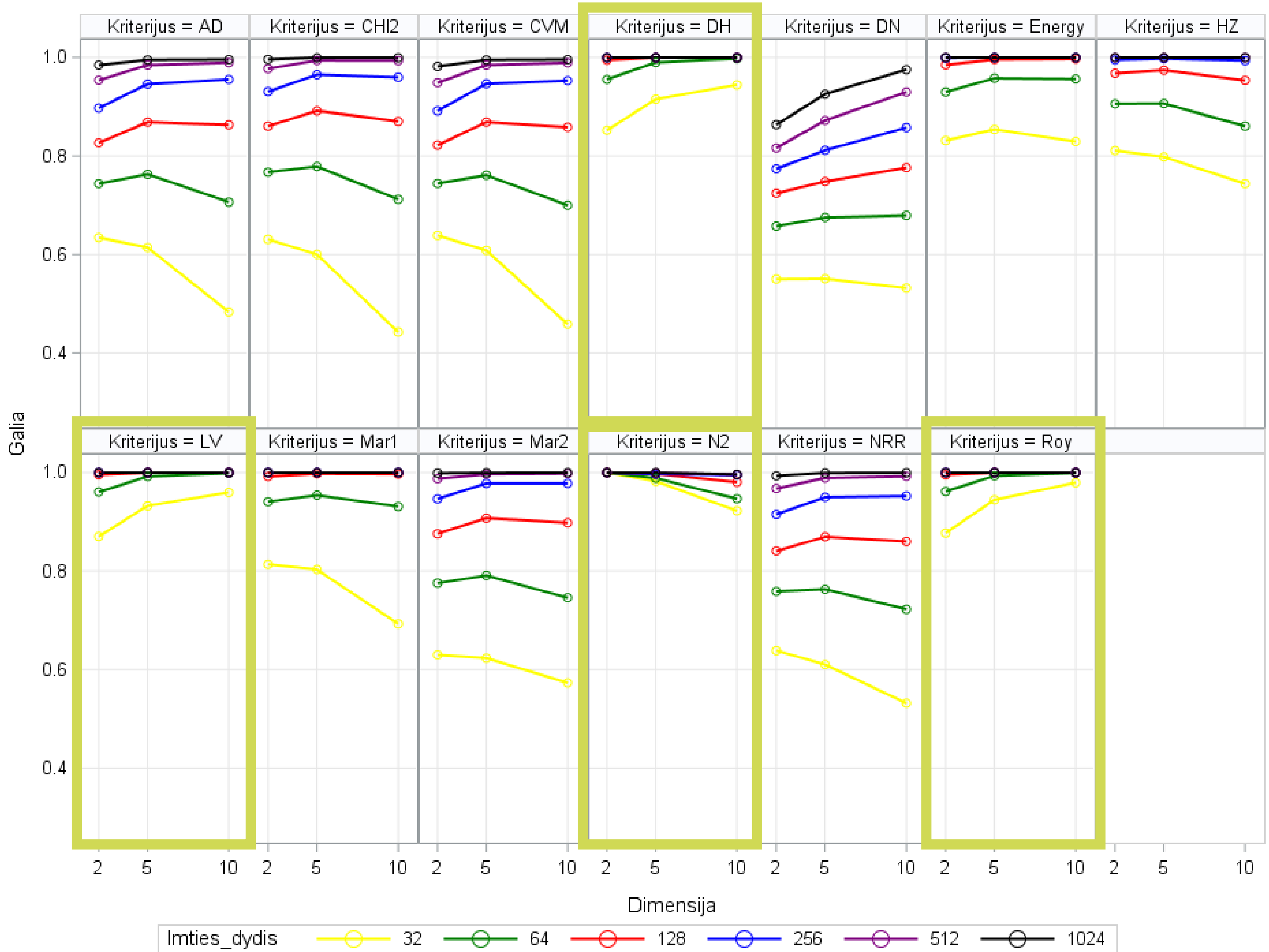
Tyrimas atliktas taikant paskirstytuosius skaičiavimus su:

- **4** skirstinių grupėmis;
- **6** imčių tūriais $(32,64,128,256,512,1024)$;
- **13** kriterijų (įskaitant ir naujai sukonstruotą);
- **3** dimensijos ($d = 2$, $d = 5$ ir $d = 10$);
- **1 000 000** nepriklausomų imčių.

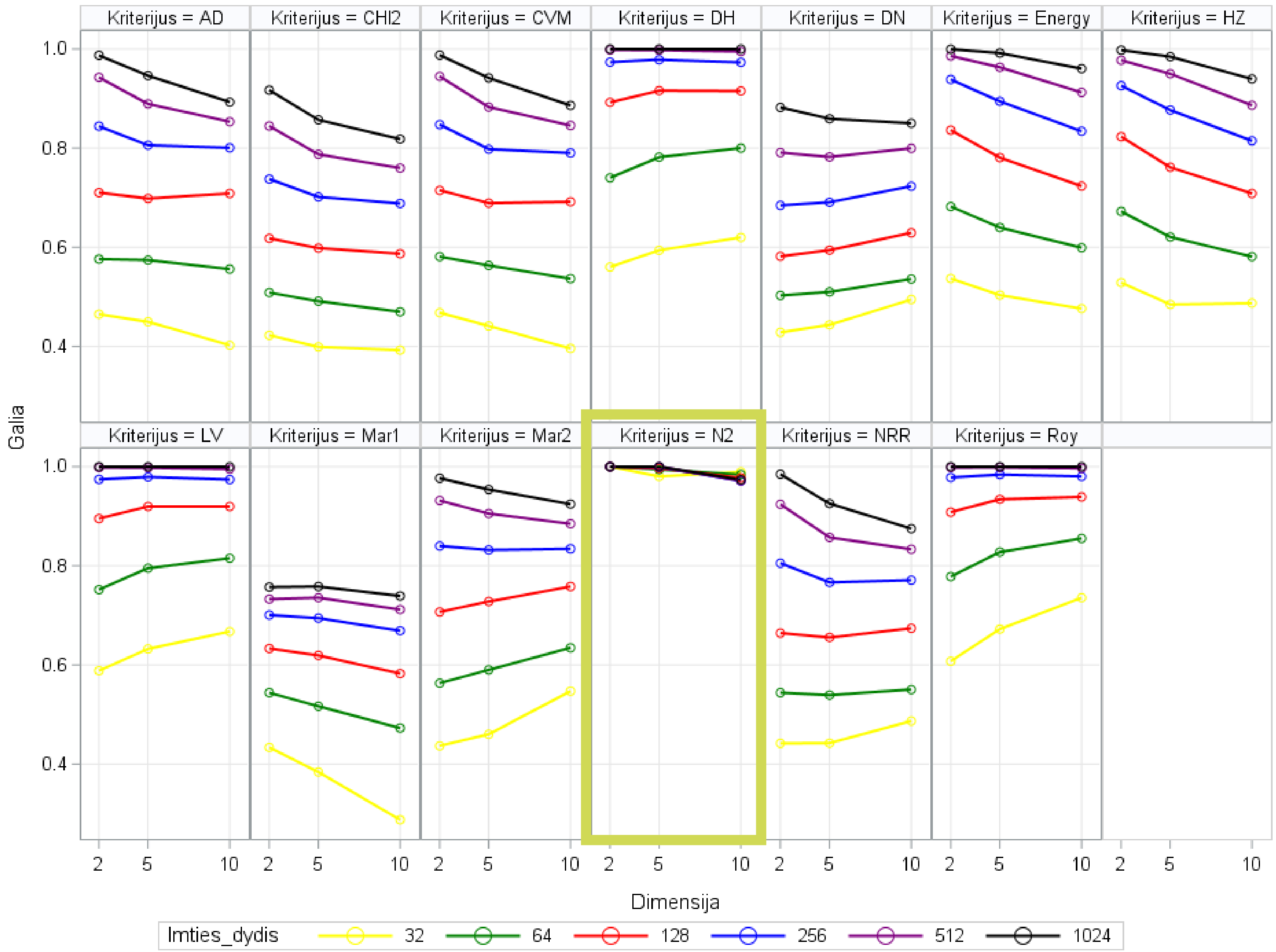
Simetrinių skirstinių grupė



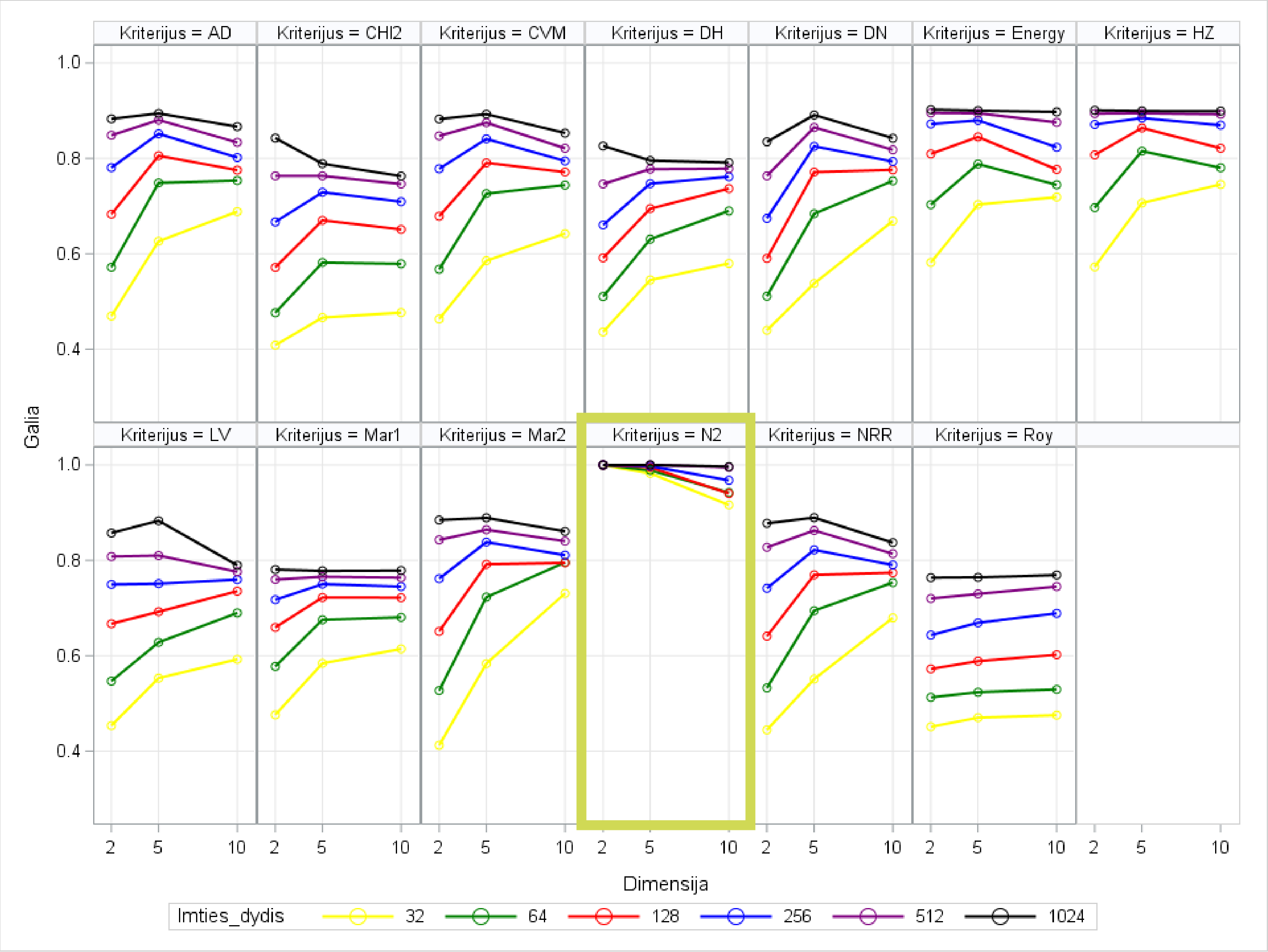
Asimetrinių skirstinių grupė



Mišriųjų skirstinių grupė



Normaliojo skirstinio mišinių grupė



Išvados ir rekomendacija

- Remiantis N-metrikos teorija, pasiūlytas suderinamumo hipotezės tikrinimo kriterijus, kuris yra galingas kitų populiarių vienmačių kriterijų konkurentas. Darbe parodyta, kad pasiūlytas kriterijus tinkamas naudoti simetriniams duomenų rinkiniams, kurių imties dydis $n \geq 112$, asimetriniams duomenų rinkiniams, kurių imties dydis $n \geq 118$, ir modifikuotų normaliųjų skirstinių duomenų rinkiniams, kurių imties dydis yra $n \geq 88$.
- Atliktas naujos branduolio funkcijos parametrų kalibravimas leido parinkti tokią jos formą, kuri vertinimo poslinkį, taikant pasiūlytą N-metrika paremtą suderinamumo hipotezės tikrinimo kriterijų, sumažino iki artimo optimaliam.
- Sukonstruotas pasiskirstymo tankių skirtumo vertinimu ir apvertimo formulės taikymu paremtas, daugiamatis suderinamumo hipotezės tikrinimo kriterijus. Daugiamačių suderinamumo hipotezės tikrinimo kriterijų palyginimas parodė, kad sukonstruotą kriterijų saikingo dydžio imtims geriausia naudoti 2-mačiams duomenims. Didesnės dimensijos duomenims šį kriterijų rekomenduojama naudoti su dideliais imties dydžiais ($n \geq 512$).
- Parodyta, kad daugiamačiai suderinamumo hipotezės tikrinimo kriterijai – Chi-kvadrato, Doorniko ir Hanseno, Energy, Henzes ir Zirklerio, Lobato ir Velasco, Nikulino, Rao ir Robsono, Roystono bei naujai pristatytas – yra galingiausi šiame tyrime naudotų realių duomenų pasiskirstymų atvejais.

2022 m. vasario 15 d.

Doktorantė: Jurgita Arnastauskaitė

Vadovas: doc. dr. Tomas Ruzgas