Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Karmeli Kaasik

# Energiakauguse rakendamine normaaljaotuse eelduse testimisel

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Märt Möls

Tartu 2024

**ENERGIAKAUGUSE RAKENDAMINE NORMAALJAOTUSE EELDUSE TESTIMISEL**

Bakalaureusetöö

Karmeli Kaasik

**Lühikokkuvõte**

Käesoleva töö eesmärgiks on selgitada energiakauguse olemust ning kasutamisvõimalusi ja uurida, millistes olukordades aitaks energiastatistiku kasutamine statistilise analüüsi tulemusi paremaks muuta. Töö esimeses kolmes peatükis tuuakse välja energiakauguse mõiste ja demonstreeritakse selle kasutamist jaotuste testimisel. Neljandas peatükis võrreldakse energiatesti teiste normaaljaotuse testidega nii ühe kui ka mitmemõõtmelisl juhul. Viies peatükk keskendub energiatesti kasutamisele struktuurivõrrandite mudelite eelduste kontrollimisel. Energiatest osutub heaks alternatiiviks teistele normaaljaotuste testidele ning selle kasutamine aitab struktuurivõrrandite mudeli tehtavate vigade arvu vähendada.

**CERCS teaduseriala:** P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

**Märksõnad:** Energiastatistik, normaaljaotuse testid, simulatsioonid, struktuurivõrrandite mudelid

**APPLICATIONS OF ENERGY DISTANCE IN NORMALITY TESTING**

Bachelor thesis

Karmeli Kaasik

**Abstract**

The aim of this bachelor's thesis is to explore the application of energy distance in statistical analyses and find out in which situations can it improve

the results of the analysis. In the first three chapters the energy distance definition is brought out along with an explanation of how to use energy statistics in distributional tests. Fourth chapter is dedicated to comparing energy test with other alternatives in both univariate and multivariate normality testing. The last chapter focuses on the use of energy statisic within structural equation models. The energy test appears to be a relevant alternative to other normality tests as well as reduce the number of errors in structural equation models.

**CERCS research specialisation:** P160 Statistics, operations research, programming, financial and actuarial mathematics.

**Key Words:** Energy statistic, tests of normality, siumulations, structural equations models

# Sisukord

**Sissejuhatus** **5**

**1 Energiakaugus** **6**

1.1 Energiakauguse valemi üldkuju . . . . . 6

1.1.1 Näide energiakauguse arvutamisest . . . . . 7

1.2 Energiastatistik jaotuse testimiseks . . . . . 8

1.2.1 Bernoulli jaotusse kuulumise testimine . . . . . 9

**2 Ühemõõtmelisele normaaljaotusele vastavuse testimine** **11**

**3 Mitmemõõtmelise normaaljaotuse testimine** **13**

**4 Testide võrdlus** **15**

4.1 Ühemõõtmeline juht . . . . . 15

4.1.1 Andmete pärinemine $t$-jaotusest . . . . . 15

4.1.2 Andmed normaaljaotuste segujaotusest . . . . . 16

4.2 Mitmemõõtmeline juht . . . . . 19

4.2.1 Kahemõõtmeline andmestik . . . . . 19

4.2.2 Kaheksamõõtmeline andmestik . . . . . 21

**5 Energiatest struktuurivõrrandites** **24**

**Kokkuvõte** **26**

**Kasutatud allikad** **28**

**Lisa 1. Monte Carlo meetodil Bernoulli jaotuse testimine** **30**

**Lisa 2. Normaaljaotuse testimine, kui andmed pärinevad t-jaotusest 31**

**Lisa 3. Normaaljaotuse testimine juhul, kui andmed pärinevad normaaljaotuste segujaotusest 33**

**Lisa 4. Normaaljaotuse testimine kahemõõtmelisel mitte-normaaljaotusest andmestikul 35**

**Lisa 5. Normaaljaotuse testimine kaheksamõõtmelisel andmestikul 37**

**Lisa 6. Energiatesti kasutamine struktuurivõrrandites 39**

# Sissejuhatus

Paljude statistiliste meetodite ja mudelite eelduseks on kas andmete või jääkide kuulumine teatud jaotuste peresse, milleks on sageli normaaljaotus. Testimaks, kas andmed tõesti normaaljaotusest pärinevad, on mitmeid viise. Ühemõõtmelisel juhul on tuntuimateks Shapiro–Wilki ja Kolmogorov–Smirnovi testid, kuid mitmemõõtmelisel juhul saab kasutada näiteks Doornik-Hanseni või Henze-Zirkleri teste. Lisaks eelnevale on võimalik andmete kuulumist kindlasse jaotusse kontrollida ka andmete energiakauguse ning sellel põhineva energiatesti abil.

Käesolevas töös võrreldaksegi energiakaugusel baseeruvat normaaljaotuse testi teiste eksisteerivate alternatiivsete meetoditega nii ühe- kui ka mitmemõõtmelisel juhul erinevate jaotuste tuvastamisel. Töö eesmärkideks on kirjeldada energiastatistiku olemust ja kasutusvõimalusi ning selgitada välja, kas energiakauguse kasutamine aitab statistilise analüüsi tulemusi paremaks muuta ja millistes olukordades võiks seda eelistada teistele meetoditele.

Töö jaguneb viieks peatükiks. Esimeses peatükiks tuuakse välja, mida energiakaugus endast kujutab ning kuidas saab seda rakendada jaotuste testimisel. Samuti on selles peatükis toodud näited energiakauguse ja -statistiku arvutamisest. Teises ning kolmandas peatükis uuritakse, milliseks muutub energiastatistiku kuju, kui soovime testida ühe- või mitmemõõtmelist normaaljaotust. Mõlemal juhul on välja toodud ka algoritmid Monte Carlo meetodil normaaljaotuse testimiseks juhul kui oletatava normaaljaotuse parameetrid tuleb hinnata valimi põhjal.

Neljas peatükk keskendub energiatesti võrdlusele teiste tuntud normaaljaotuse testidega. Testide võimsuste võrdlused viidi läbi nii ühe- kui ka mitmemõõtmelisel juhul erinevate valimimahtude korral. Ühemõõtmelisel juhul uuriti testide käitumist, kui andmed pärinesid tegelikult $t$-jaotusest või normaaljaotuste segujaotusest, mitmemõõtmisel juhul vaadeldi kahe- ning kaheksamõõtmelist mitte-normaaljaotust. Viiendas peatükis vaadatakse, kas ja kuidas aitab energiatesti kasutamine kaasa struktuurivõrrandite mudelite eelduste kontrollimisel.

# 1 Energiakaugus

Andmete vahelist kaugust, mis baseerub Newtoni gravitatsiooni potentsiaalse energia printsiibil, nimetatakse andmete energiaks. Andmevaatlustest mõeldakse kui kehadest, mille potentsiaalne energia on null siis ja ainult siis, kui aluseks olev statistiline nullhüpotees kehtib. Andmete energia on vaatluste vaheliste kauguste funktsioon. (Székely ja Rizzo, 2023, lk 4)

Paljudel klassikalistel statistilistel testidel ja meetoditel on energiakaugusel põhinevad vasted. Andmeenergia abil saab leida nii tunnuste sõltuvust kui ka homogeensust, seda saab rakendada dispersioonanalüüsi ja klasteranalüüsi korral ning energiastatistiku abil saab testida valimi erinevatesse jaotustesse kuulumist. (Székely ja Rizzo, 2017)

Edaspidises vaadatakse vaid Eukleidilisi ehk ruutkaugusi. Olgu antud kaks punkti kahemõõtmelises ruumis, $x = (x_1, x_2)$ ja $y = (y_1, y_2)$ Eukleidisise kauguse arvutusvalem nende punktide vahel on $|x-y| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. Rohkem kui kahemõõtmelises ruumis jääb valemi põhimõte samaks. Valem saavutab kuju

$$\sqrt{(x_1 - y_1)^2 + \ldots + (x_d - y_d)^2},$$

mille abil saamegi leida $d$-mõõtmelises ruumis punktide $(x_1, \ldots, x_d)$ ja $(y_1, \ldots, y_d)$ vahelise kauguse. (Tabak, 2014, lk 150)

## 1.1 Energiakauguse valemi üldkuju

Olgu $X \in \mathbb{R}^d$ ja $Y \in \mathbb{R}^d$ sõltumatud juhuslikud suurused ning $X'$ ja $Y'$ neile vastavad identselt jaotatud sõltumatud koopiad. $X$ ja $Y$ vaheline kaugus on sel juhul defineeritud kui

$$\mathcal{E}(X, Y) = 2\mathbb{E}|X - Y|_d - \mathbb{E}|X - X'|_d - \mathbb{E}|Y - Y'|_d, \tag{1}$$

kus $\mathbb{E}$ tähistab keskväärtust, $|\cdot|_d$ Eukleidilist kaugust $d$-mõõtmelises ruumis ning kus kehtib $\mathbb{E}|X|_d < \infty$ ja $\mathbb{E}|Y|_d < \infty$. Selgub, et alati $\mathcal{E} \geq 0$ ning $\mathcal{E} = 0$ siis ja ainult siis, kui $X$ ja $Y$ jaotused ühtivad. (Székely ja Rizzo, 2013)

### 1.1.1 Näide energiakauguse arvutamisest

Olgu $X$ ja $Y$ Bernoulli jaotusega sõltumatud juhuslikud suurused, parameetritega vastavalt $p_1$ ja $p_2$. Sellisel juhul

$$\begin{aligned}\mathbb{E}|X-Y| &= \sum_{x=0}^{1}\sum_{y=0}^{1}|x-y|\cdot P(X=x)\cdot P(Y=y)\\ &= 1\cdot P(X=1)\cdot P(Y=0)+1\cdot P(X=0)\cdot P(Y=1)\\ &= p_1\cdot(1-p_2)+(1-p_1)\cdot p_2.\end{aligned}$$

Samamoodi saab leida $\mathbb{E}|X-X'| = 2\cdot p_1\cdot(1-p_1)$ ja $\mathbb{E}|Y-Y'| = 2\cdot p_2\cdot(1-p_2)$. Järelikult kasutades valemit (1), saab arvutada energiakauguse kahe Bernoulli jaotusega juhusliku suuruse vahel järgnevalt:

$$\begin{aligned}\mathcal{E}(X,Y) &= 2\cdot(p_1\cdot(1-p_2)+(1-p_1)\cdot p_2)-2\cdot p_1\cdot(1-p_1)-2\cdot p_2\cdot(1-p_2)\\ &= 2\cdot p_1^2+2\cdot p_2^2-4\cdot p_1p_2 = 2\cdot(p_1-p_2)^2.\end{aligned} \tag{2}$$

Asendades valemisse (2) parameetrite $p_1$ ja $p_2$ kindlad väärtused, saamegi leida kahe konkreetse jaotuse energiakauguse. Näeme ka, et kui $p_1 = p_2$, lihtsustub see valem kujule

$$\mathcal{E}(X,Y) = 2\cdot(p_1-p_1)^2 = 0,$$

mis tähendab, et kahe sama parameetriga Bernoulli jaotusega juhusliku suuruse energiakaugus on 0.

## 1.2 Energiastatistik jaotuse testimiseks

Olgu $\mathbb{X} = (X_1, \ldots, X_n)$ $n$-liikmeline juhuslik valim jaotusega $F$ ning $x_1, \ldots, x_n$ selle valimi väärtused. Testimaks hüpoteese $H_0 : F = F_0$ ja $H_1 : F \neq F_0$, ehk teisisõnu testimaks valimi jaotuse vastavust oletatavale jaotusele $F_0$, saab kasutada ühe valimi põhist energiastatistikut. Statistiku kuju on järgnev:

$$\mathcal{E}_n(\mathbb{X}, F_0) = n \left( \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}|x_i - X| - \mathbb{E}|X - X'| - \frac{1}{n^2} \sum_{l=1}^{n} \sum_{m=1}^{n} |x_l - x_m| \right), \tag{3}$$

kus $X$ ning $X'$ on sõltumatud jaotusega $F_0$ juhuslikud suurused. (Székely ja Rizzo, 2013)

Ainult statistiku väärtuse põhjal ei saa aga veel teha järeldusi valimi jaotusse kuulumise kohta. Hüpoteeside $H_0 : F = F_0$ ja $H_1 : F \neq F_0$ testimiseks kasutatakse enamasti Monte Carlo simulatsioone. Székely ja Rizzo (2023, lk 53) on oma raamatus välja toonud lihtsa algoritmi simulatsioonide põhjal valimi hüpoteetilisse jaotusse kuulumise testimiseks. Valides korduste arvuks $R$ ning olulisuse nivooks $\alpha$ on algoritm järgnev:

1. Arvutada vaadeldud valimi $\mathbb{X}$ korral statistiku väärtus $T_0 = \mathcal{E}_n(\mathbb{X}, F_0)$ valemi (3) abil.
2. Korrata järgnevaid samme $j = 1, \ldots, R$ korda:
   (a) Genereerida valim $\mathbb{X}^*$ jaotusest $F_0$.
   (b) Arvutada valimile $\mathbb{X}^*$ vastav teststatistik $T^{(j)} = \mathcal{E}_n(\mathbb{X}^*, F_0)$ valemi (3) abil.
3. Leida $1 - \alpha$ kvantiil väärtuste $T^{(1)}, \ldots, T^{(R)}$ hulgast.
4. Kui $T_0$ on suurem või võrdne leitud kvantiiliga, võtta vastu alternatiivne hüpotees, vastasel juhul jääda nullhüpoteesi juurde.

### 1.2.1 Bernoulli jaotusse kuulumise testimine

Olgu $X$ Bernoulli jaotusega parameetriga $p = 0{,}1$ juhuslik suurus ning olgu antud kolmeelemendiline valim $\mathbb{X} = (X_1, X_2, X_3)$ väärtustega $x_1 = 0, x_2 = 0, x_3 = 1$. Soovime testida, kas eeltoodud valim võiks kuuluda samasse jaotusesse, kuhu kuulub $X$ ehk kas valim võiks olla võetud Bernoulli jaotusesest parameetriga 0,1. Leiame statistiku väärtuse valemi (3) abil:

$$\begin{aligned}
\mathcal{E}_n &= n\left(\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}|x_i - X| - \mathbb{E}|X - X'| - \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n}|x_l - x_m|\right) \\
&= 3\left(\frac{2}{3}\sum_{i=1}^{3}(x_i \cdot 0{,}9 + |x_i - 1| \cdot 0{,}1) - 2 \cdot 0{,}1 \cdot 0{,}9 - \frac{1}{3^2}\sum_{l=1}^{3}\sum_{m=1}^{3}|x_l - x_m|\right) \\
&= 3\left(\frac{2}{3} \cdot 1{,}1 - 0{,}18 - \frac{1}{9} \cdot 4\right) = \frac{49}{150}.
\end{aligned}$$

Kuna antud näites on valimimaht väike, saame teststatistiku jaotuse ka analüütiliselt välja arvutada. Kasutame standardset olulisuse nivood 0,05. Leiame kõikide $x_1, x_2$ ja $x_3$ väärtuste kombinatsioonide korral $\mathcal{E}_n$ väärtused. Kombinatsioone on kokku 8, leiame nende kõigi korral energiastatistiku väärtused ja tõenäosused, et selline kombinatsioon esineb. Tulemused on välja toodud tabelis 1.

Tabel 1: Energiastatistiku väärtused kolmeelemendiliste valimite korral, kus vaatlused on Bernoulli jaotusest

| $x_1$ | $x_2$ | $x_3$ | $P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ | $\mathcal{E}_n$ |
|---|---|---|---|---|
| 0 | 0 | 0 | $0{,}9 \cdot 0{,}9 \cdot 0{,}9 = 0{,}729$ | $\frac{3}{50} = 0{,}0600$ |
| 1 | 0 | 0 | $0{,}1 \cdot 0{,}9 \cdot 0{,}9 = 0{,}081$ | $\frac{49}{150} = 0{,}3267$ |
| 0 | 1 | 0 | $0{,}9 \cdot 0{,}1 \cdot 0{,}9 = 0{,}081$ | $\frac{49}{150} = 0{,}3267$ |
| 0 | 0 | 1 | $0{,}9 \cdot 0{,}9 \cdot 0{,}1 = 0{,}081$ | $\frac{49}{150} = 0{,}3267$ |
| 1 | 1 | 0 | $0{,}1 \cdot 0{,}1 \cdot 0{,}9 = 0{,}009$ | $\frac{289}{150} = 1{,}9267$ |
| 1 | 0 | 1 | $0{,}1 \cdot 0{,}9 \cdot 0{,}1 = 0{,}009$ | $\frac{289}{150} = 1{,}9267$ |
| 0 | 1 | 1 | $0{,}9 \cdot 0{,}1 \cdot 0{,}1 = 0{,}009$ | $\frac{289}{150} = 1{,}9267$ |
| 1 | 1 | 1 | $0{,}1 \cdot 0{,}1 \cdot 0{,}1 = 0{,}001$ | $\frac{243}{50} = 4{,}8600$ |

Tabeli põhjal saame leida olulisuse tõenäosuse, mille väärtuseks on antud juhul $p = 3 \cdot 0{,}081 + 3 \cdot 0{,}009 + 0{,}001 = 0{,}271$. Peame jääma nullhüpoteesi juurde - andmed võivad olla Bernoulli jaotusest parameetriga 0,1. Kui testida jaotusesse kuuluvust Monte Carlo meetoditega, kasutades $R = 1000$ simulatsiooni, saame olulisuse tõenäosuseks 0,277. Meetodi rakendamine R-is (R Core Team, 2021) on välja toodud lisas 1. Sellisel meetodil testides saame samamoodi jääda nullhüpoteesi juurde ja öelda, et valim tuleneb välja pakutud jaotusest. Mõlemal juhul tulid juba väikese valimimahu korral $p$-väärtused sarnased, seega töötab algoritm 1 üsna hästi.

# 2 Ühemõõtmelisele normaaljaotusele vastavuse testimine

Soovides testida, kas valim $\mathbb{X} = (X_1, \ldots, X_n)$ väärtustega $x_1, \ldots, x_n$ võiks kuuluda ühemõõtmelisse normaaljaotusesse keskväärtusega $\mu$ ja standardhälbega $\sigma$, muutuvad valemi (3) liikmed $\mathbb{E}|X - X'|$ ja $\mathbb{E}|x_i - X|$.

Liikme $\mathbb{E}|X - X'|$ saab leida kui $\mathbb{E}|Y|$, kus $Y = X - X' \sim \mathcal{N}(0, 2\sigma^2)$. Seega, tuginedes tingliku keskväärtuse omadustele,

$$\begin{aligned}
\mathbb{E}|Y| &= \mathbb{E}(Y|Y > 0) \cdot P(Y > 0) + \mathbb{E}(Y|Y < 0) \cdot P(Y < 0) = \mathbb{E}(Y|Y > 0) \\
&= \left( \int_0^\infty y \cdot \frac{1}{2\sigma\sqrt{\pi}} \exp\left( \frac{-y^2}{4\sigma^2} \right) dy \right) \Big/ 0{,}5 = \frac{1}{\sigma\sqrt{\pi}} \int_0^\infty y \cdot \exp\left( \frac{-y^2}{4\sigma^2} \right) dy \\
&= \frac{-2\sigma^2}{\sigma\sqrt{\pi}} \int_0^{-\infty} \exp(u)\, du = \frac{-2\sigma}{\sqrt{\pi}} \left( \exp(u) \Big|_0^{-\infty} \right) = \frac{-2\sigma}{\sqrt{\pi}} \cdot (-1) = \frac{2\sigma}{\sqrt{\pi}}.
\end{aligned}$$

Teine liige muutub ühemõõtmelist normaaljaotust testides järgnevalt:

$$\mathbb{E}|x_i - X| = 2(x_i - \mu)F(x_i) + 2\sigma^2 f(x_i) - (x_i - \mu),$$

kus $F(x)$ ja $f(x)$ on nullhüpoteesile vastava normaaljaotuse jaotus- ja tihedusfunktsioon (Székely ja Rizzo, 2013).

Seega, asendades eeltoodud väärtused valemisse (3), saamegi energiastatistiku ühemõõtmelise normaaljaotuse testimiseks:

$$\begin{aligned}
\mathcal{E}_n(\mathbb{X}) = n\Bigg( &\frac{2}{n} \sum_{i=1}^{n} [2(x_i - \mu)F(x_i) + 2\sigma^2 f(x_i) - (x_i - \mu)] - \\
&\frac{2\sigma}{\sqrt{\pi}} - \frac{1}{n^2} \sum_{l=1}^{n} \sum_{m=1}^{n} |x_l - x_m| \Bigg).
\end{aligned} \tag{4}$$

Kui valim eelnevalt standardiseerida, kasutades valimi keskmist ning valimi standardhälvet, saab normaaljaotust testida endiselt valemi (4) abil, kuid sellisel ju-

hul tähistab $F(x)$ standardse normaaljaotuse jaotusfunktisooni ning $f(x)$ tihedusfunktsiooni (Székely ja Rizzo, 2023, lk 54). Sellest tulenevalt saabki sarnaselt algoritmile 1 kirja panna algoritmi Monte Carlo meetodil normaaljaotuse testimiseks, kui hüpoteetilise jaotuse keskväärtus ja standardhälve on teadmata (Székely ja Rizzo, 2023, lk 56):

1. Leida valimi $\mathbb{X}$ keskmine $\overline{x}$ ning valimi standardhälve $s$.
2. Moodustada valim $\mathbb{Y}$, standardiseerides selleks valimi $\mathbb{X}$ väärtused: $y_i = \frac{x_i - \overline{x}}{s}$
3. Arvutada valimile $\mathbb{Y}$ vastav statistiku väärtus $T_0 = \mathcal{E}_n(\mathbb{Y})$ valemi (4) abil, kus $F(x)$ ja $f(x)$ on vastavalt standardse normaaljaotuse jaotus- ja tihedusfunktisoon.
4. Korrata järgnevaid samme $j = 1, \ldots, R$ korda:
   (a) Genereerida valim $\mathbb{X}^*$ jaotusest $N(0, 1)$.
   (b) Leida valimi $\mathbb{X}^*$ keskväärtus $\overline{x^*}$ ja standardhälve $s^*$.
   (c) Moodustada valim $\mathbb{Y}^*$, standardiseerides valimi $\mathbb{X}^*$ väärtused.
   (d) Arvutada standardiseeritud valimile vastav teststatistik $T^{(j)} = \mathcal{E}_n(\mathbb{Y}^*)$ valemi (4) abil.
5. Leida $1 - \alpha$ kvantiil väärtuste $T^{(1)}, \ldots, T^{(R)}$ hulgast.
6. Võtta vastu alternatiivne hüpotees kui $T_0$ on suurem või võrdne leitud kvantiiliga, vastasel juhul jääda nullhüpoteesi juurde.

Standaridseerides muutuvad küll valimi elemendid omavahel sõltuvaks, kuid Monte Carlo simulatsiooni korral võrreldakse sõltuvat valimit nullhüpoteesi korral saadud samuti sõltuvate valimitega, mistõttu saab simulatsioonis valemit (4) edukalt kasutada. Samasugust algoritmi on energiastatistiku arvutamiseks kasutatud ka R-i paketis *energy* (Rizzo ja Szekely, 2022).

# 3 Mitmemõõtmelise normaaljaotuse testimine

*Peatükk on kirjutatud Székely ja Rizzo, 2005 ning Székely ja Rizzo, 2023 põhjal.*

Selleks, et testida energiatestiga, kas valim $\mathbb{X} = (X_1, \ldots, X_n)$ võiks kuuluda $d$-mõõtmelisse normaaljaotusesse keskväärtuste vektoriga $\mu$ ning kovariatsioonimaatriksiga $\Sigma$, tuleb esmalt valim transformeerida. Iga $X_i$, $i \in \{1, \ldots n\}$ korral antud valimist leiame vastavad väärtused $Y_i = \Sigma^{-1/2}(X_i - \mu)$, et vaatlused oleksid samast jaotusest kui $Z \sim N_d(\mu = 0, \Sigma = I_d)$. Sellisel juhul muutuvad valemi (3) liikmed:

$$\mathbb{E}|Y_i - Z| = \frac{\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} + \sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!2^k} \frac{|Y_i|_d{}^{2k+2}}{(2k+1)(2k+2)} \frac{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(k+\frac{3}{2}\right)}{\Gamma\left(k+\frac{d}{2}+1\right)},$$

mille saab kirja panna ka hüpergeomeetrilise funktsiooni ${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{a^{\bar{n}}}{b^{\bar{n}}} \frac{z^n}{n!}$, kus $a^{\bar{n}} = \frac{\Gamma(n+a)}{\Gamma(n)}$ abil ehk järgmiselt:

$$\mathbb{E}|Y_i - Z| = \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} {}_1F_1\left(-\frac{1}{2}, \frac{d}{2}; \frac{-|Y_i|_d^2}{2}\right)$$

ning

$$\mathbb{E}|Z - Z'| = 2\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

Olgu $\kappa_d = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}$ ning ${}_1F_1$ eeltoodud hüpergeomeertiline funktsioon. Siis saame energiastatistiku lõpliku kuju mitmemõõtmelise normaaljaotuse testimiseks kirja panna järgnevalt:

$$\mathcal{E}_{n,d}(\mathbb{Y}) = n\left[\sqrt{2}\kappa_d \cdot \frac{2}{n}\sum_{i=1}^{n} {}_1F_1\left(-\frac{1}{2}, \frac{d}{2}; \frac{-|Y_i|_d^2}{2}\right) - 2\kappa_d - \frac{1}{n^2}\sum_{i,j=1}^{n} |Y_i - Y_j|\right]. \quad (5)$$

Praktikas on tavaliselt oletatava normaaljaotuse parameetrid teadmata, mistõttu tuleb need leida valimi põhjal. Olgu $X_i$, $i \in \{1, \ldots, n\}$ antud valimi $\mathbb{X}$ vektorid.

Mitmemõõtmelise normaaljaotuse testimise algoritm Monte Carlo meetodil, mis on realiseeritud ka R-i paketis *energy*, on järgnev:

1. Leida valimi keskmiste vektor $\overline{X}$ ning valimi kovaratsioonimaatriks $S = (n-1)^{-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(X_i - \bar{X}\right)^\top$.

2. Standardiseerida valimi vektorid: $Y_i = S^{-1/2}(X_i - \overline{X})$, kus $S^{-1/2}$ on maatriksi $S$ pöördmaatriks astmes 0,5.

3. Arvutada statistiku väärtus $T_0 = \mathcal{E}_{n,d}(\mathbb{Y})$ valemi (5) abil

4. Korrata järgnevaid samme $j = 1, \ldots, R$ korda:

   (a) Genereerida valim $\mathbb{X}^* = (X_1^*, \ldots, X_n^*)$ mitmemõõtmelisest normaaljaotusest $N_d(\mu = 0, \Sigma = I_d)$, kus $I_d$ on ühikmaatriks.

   (b) Leida valimi $\mathbb{X}^*$ keskväärtuste vektor $\overline{X^*}$ ja kovariatsioonimaatriks $S^*$.

   (c) Genereerida standardiseeritud valim $\mathbb{Y}^*$, kus $Y_i = S^{-1/2}(X_i - \overline{X^*})$

   (d) Arvutada standardiseeritud valimile vastav teststatistik $T^{(j)} = \mathcal{E}_{n,d}(\mathbb{Y}^*)$ valemi (5) abil.

5. Leida $1 - \alpha$ kvantiil väärtuste $T^{(1)}, \ldots, T^{(R)}$ hulgast.

6. Võtta vastu alternatiivne hüpotees kui $T_0$ on suurem või võrdne leitud kvantiiliga, vastasel juhul jääda nullhüpoteesi juurde.

# 4 Testide võrdlus

Bakalaureusetöö teises pooles võrreldakse erinevaid normaaljaotuse teste energiatestiga. Simulatsioonid viidi läbi normaaljaotusest mittetulenevate valimite korral nii ühemõõtmelisel kui ka mitmemõõtmelisel juhul. Genereeritud andmete põhjal hinnati testi võimsust erinevatel valimimahtudel. Kõikide testide korral valiti olulisuse nivooks $\alpha = 0{,}05$. Andmete genereerimine ja arvutused viidi läbi kasutades tarkvara *R* ja on välja toodud töö lisades. Joonised on koostatud paketi *ggplot2* (Wickham, 2016) abil.

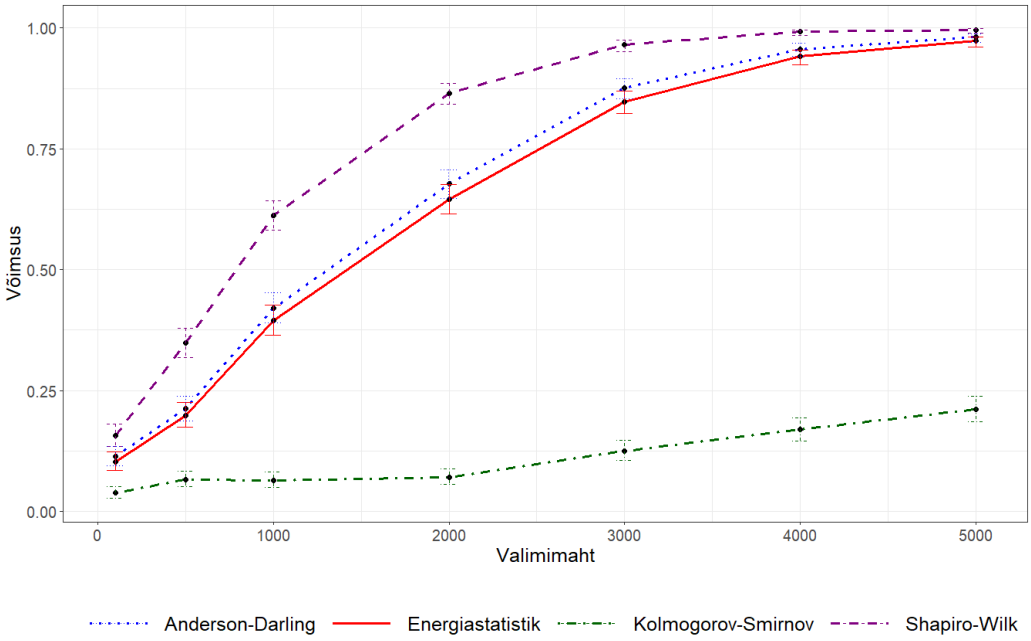
## 4.1 Ühemõõtmeline juht

Testimaks ühemõõtmelise energiastatistiku headust, genereeriti andmeid nii normaaljaotuste segujaotusest kui ka $t$-jaotusest. Testiti erinevate valimimahtude korral ning iga valimi suuruse kohta moodustati 1000 erinevat valimit alternatiivsest jaotusest, mille põhjal arvutati välja testi võimsus. Energiatesti võrreldi kolme tuntud normaaljaotuse testiga - Kolmogorov-Smirnovi, Shapiro-Wilki ning Anderson-Darlingi testidega.

### 4.1.1 Andmete pärinemine $t$-jaotusest

$T$-jaotuse puhul oli vabadusastmete arvuks valitud 15, kuna sellise parameetriga $t$-jaotus on piisavalt sarnane standardsele normaaljaotusele, kuid siiski tavapäraste valimimahtude piires tuvastatav. Tulemused fikseeriti 7 erineva valimimahu korral ehk $n \in \{100; 500; 1000; 2000; 3000; 4000; 5000\}$. Testimise tulemused on toodud välja joonisel 1, kus on näha kõikide normaaljaotuse testide võimsused tuvastada normaaljaotusest kõrvalekallet, kui andmed pärinevad tegelikult $t$-jaotusest, samuti on lisatud ka võimsuse usalduspiirid. Andmete genereerimise programm on leitav lisas 2.

Jooniselt (vt joonis 1) näeme, et kõige paremini tuvastab $t$-jaotuse normaaljaotuse asemel Shapiro-Wilki test, mis saavutab võimsuse üle 0,8 juba valimimahu $n = 2000$ juures. Energiastatistik saavutas soovitud võimsuse valimi suuruse $n = 3000$ korral, nagu ka Anderson-Darlingi test. Kolmogorov-Smirnovi testi võimsus jäi aga kõigil varasemalt fikseeritud juhtudel alla 0,25, saavutades soovitud võimsuse alles valimimahu $n = 13000$ korral. Seega, kui andmed pärinevad tegelikult $t$-jaotusest, on energiatesti tuvastusvõime võrreldes teiste testidega piisavalt hea, jäädes alla vaid Shapiro-Wilki testile.

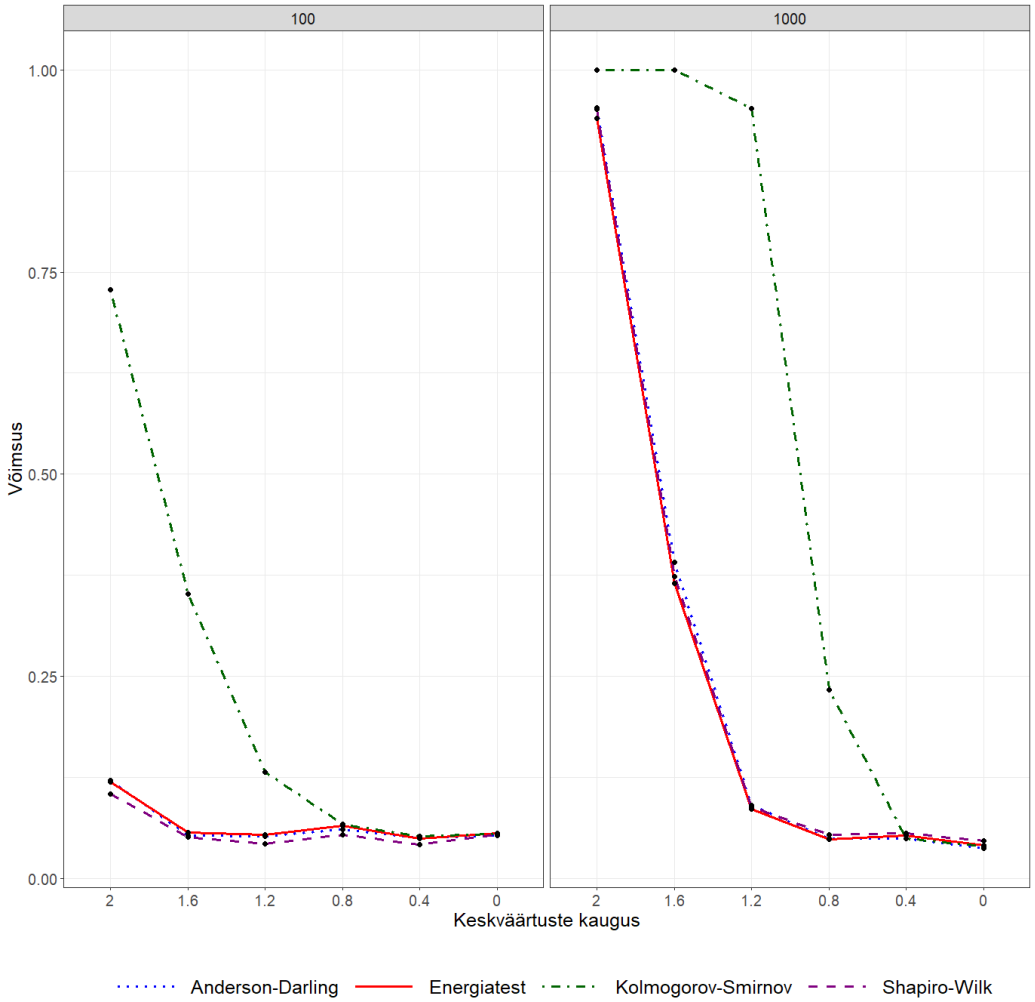



Joonis 1: Energiatesti võrdlus teiste testidega normaaljaotuse testimisel. Andmed pärinevad tegelikult $t$-jaotusest.

### 4.1.2 Andmed normaaljaotuste segujaotusest

Energiatesti võrreldi teiste testidega ka normaaljaotuste segujaotuse tuvastamisel. Segujaotus saadi valides vaatlused juhuslikult kahest normaaljaotusest, kus mõlema jaotuse saamise tõenäosus oli võrdne. Kummagi jaotuse standardhälbeks valiti $\sigma_1 = \sigma_2 = 1$, keskväärtuseid muudeti mõlemal jaotusel korraga ja vastavalt väär-

tuste hulgast $\mu_1 \in \{1; 0{,}8; 0{,}6; 0{,}4; 0{,}2; 0\}$ või $\mu_2 \in \{-1; -0{,}8; -0{,}6; -0{,}4; -0{,}2; 0\}$. Seetõttu muutus iga iteratsiooniga lähtejaotuste keskväärtuste vahe väiksemaks ja seega ka segujaotus ise standardse normaaljaotuse sarnasemaks. Samuti vaadati, kuidas testid käituvad, kui testitavaks jaotuseks on päriselt normaaljaotus ehk kui segujaotuse kahe normaaljaotuse keskväärtuste vahe on 0. Tulemused fikseeriti valimimahtude $n = 100$ ja $n = 1000$ korral.




Joonis 2: Testide võrdlus normaaljaotuse testimisel, kui andmed pärinevad normaaljaotuste segujaotusest.

Andmete genereerimise programm on leitav lisast 3 ning simulatsiooni tulemused on välja toodud joonisel 2. Joonise $y$-telg näitab testi võimsust ning $x$-telg seda, kui

kaugel on kahe normaaljaotuse keskväärtused teineteisest. Näiteks, kui $\mu_1 = 0{,}8$ ning $\mu_2 = -0{,}8$, on $x$-teljel vastav väärtus 1,6.

Nagu näha jooniselt 2, on segujaotuse tuvastamisel parim normaaljaotuse test Kolmogorov-Smirnovi test, mis saavutas kummagi valimimahu korral kõige suurema võimsuse. Eriti hästi tuvastas see test erinevust, kui segujaotuse keskväärtuste kaugus oli suurem kui 1, täpsemalt kombinatsioonide $\mu_1 = 1, \mu_2 = -1$; $\mu_1 = 0{,}8, \mu_2 = -0{,}8$ ja $\mu_1 = 0{,}6, \mu_2 = -0{,}6$ korral. Energiatest, Anderson-Darlingi ja Shapiro-Wilki testid töötasid kõik üsna sarnaselt ja märkimisväärselt väiksema võimsusega kui Kolmogorov-Smirnovi test. Kui keskväärtuste vaheline kaugus oli 0, ehk oli tegemist standardse normaaljaotusega, saavutasid kõik testid võimsuse 0,05 ümbruses, mis oli varasemalt fikseeritud olulisuse nivoo.

Eelneva põhjal saame järeldada, et energiatest on piisavalt hea alternatiiv ühemõõtmelise normaaljaotuse testimiseks. Olenevalt aga originaalsete andmete tegelikust jaotusest, võib mõni teine test sama valimimahu korral anda paremaid tulemusi ning energiatesti kasutades oleks tarvis suuremat valimimahtu, et saavutada soovitud testi võimsust.

## 4.2 Mitmemõõtmeline juht

Mitmemõõtmelise normaaljaotuse testimisel võrreldi energiatesti kolme teise tuntud testiga. Nendeks olid Henze-Zirkeli, Roystoni ja Doornik-Hanseni testid. Testide läbiviimiseks kasutati R-i paketti MVN (Korkmaz, Goksuluk ja Zararsiz, 2021), milles sisalduvad kõik vaatluse all olevad testid. Normaaljaotust testiti kahel erineval juhul - kahemõõtmelise ning kaheksamõõtmelise andmestiku korral. Kasutati endiselt tavapärast olulisuse nivood, $\alpha = 0{,}05$.

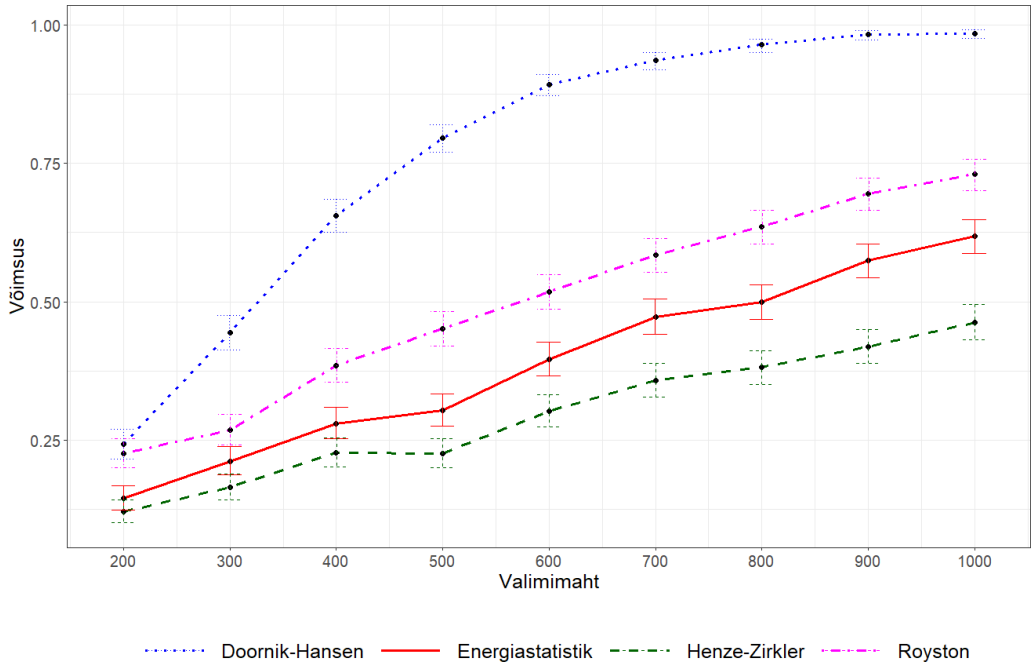
### 4.2.1 Kahemõõtmeline andmestik

Kahemõõtmelise andmestiku korral valiti üks komponent standardsest normaaljaotusest, kuid teise komponendi iga liikme $u$ jaoks rakendati järgnevat funktsiooni:

$$f(u) = |u|^a \cdot \text{sign}(u), \tag{6}$$

kus $a$ väärtuseks oli 1,1 esimesel simulatsioonil ning 1,2 teisel juhul. Mõlema simulatsiooni korral fikseeriti tulemused 9 valimimahul, $n \in \{200; 300; 400; 500; 600; 700; 800; 900; 1000\}$. Iga $n$ väärtuse põhjal koostati 1000 selles suuruses valimit, mille põhjal leiti eeltoodud testide $p$-väärtused. Nende kaudu saadi võimsused tuvastamaks, et andmed ei pärine kahemõõtmelisest normaaljaotusest. Binoomtesti abil leiti kõikidele saadud võimsustele ka usalduspiirid. Sellisel viisil andmete simuleerimise programm on välja toodud lisas 4.

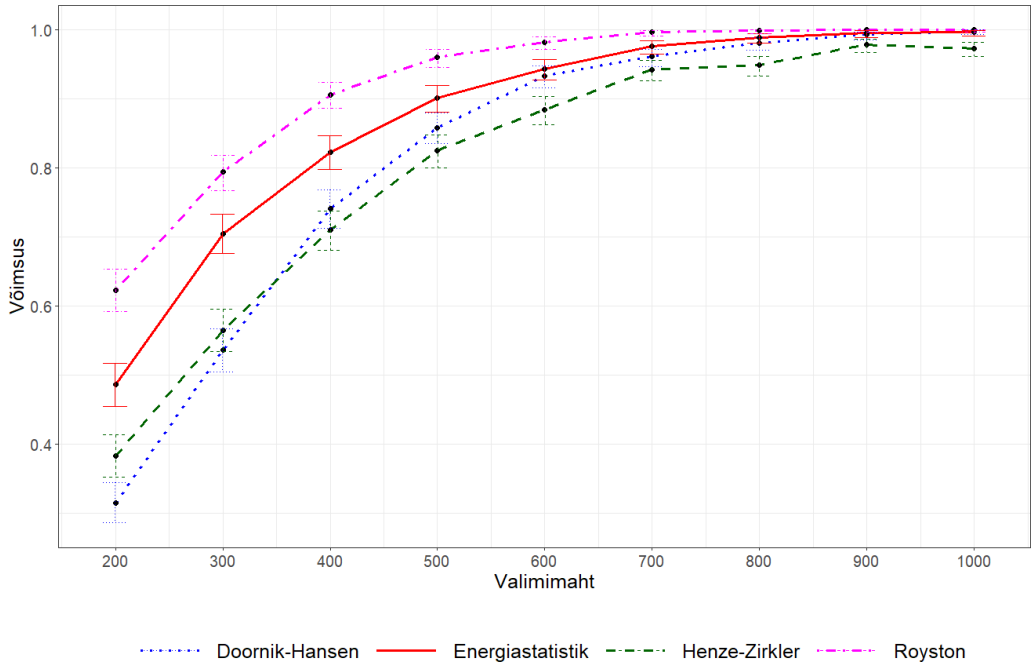
Joonisel 3 on toodud välja testide võimsused, kui kahemõõtmelise normaaljaotuse ühte komponenti on muudetud funktsiooni (6) abil, kasutades parameetrina $a = 1{,}1$. Näeme, et ainuke test, mis eelmainitud valimimahtude piires saavutab võimsuste üle 0,8, on Doornik-Hanseni test. Roystoni test suudab parimal juhul saavutada võimsuse 0,73, energiatest vaid 0,62. Henze-Zirkleri test sai parimaks võimsuseks aga ainult 0,46. Ka usaldusintervallide põhjal eristuvad testid üksteisest selgelt ning parimaks saamegi pidada Doornik-Hanseni testi.




Joonis 3: Testide võrdlus normaaljaotuse testimisel, kui andmed pärinevad kahemõõtmelisest jaotusest, mille üks komponent on normaaljaotusega, kuid teist on muudetud valemis (6) toodud funktsiooniga, parameeter $a = 1{,}1$

Testide võimsused, kui kahemõõtmelise normaaljaotuse ühele komponendile on rakendatud funktsiooni 1 parameetriga $a = 1{,}2$, on välja toodud joonisel 4. Näeme, et sellisel juhul on parimaks testiks Roystoni test, kuid ka teised ei jää sellest palju maha. Tulemused on selgelt paremad, kui eelneval juhul. Roystoni test saavutas võimsuse üle 0,8 valimimahu 300 korral, Energiatest 400, Doornik-Hanseni 500 ning Henze-Zirkeli test samuti valimi suuruse 500 korral. Suuremate valimimahtude korral ei ole aga testid üksteisest oluliselt erinevate võimsustega.

Saab järeldada, et energiatest ei pruugi olla parim meetod kahemõõtmelise normaaljaotuse tuvastamiseks, kui andmete erinevus normaaljaotusest seisneb ühe komponendi muutmises mitte-normaaljaotuseks. Sellisel juhul võiks kasutada Roystoni testi, mis sai mõlemas katses hea tulemuse. Küll aga on energiatest kindlasti piisav alternatiiv teistele testidele ning seda saab ka sellisel juhul mitmemõõtmelise normaaljaotuse testimiseks edukalt rakendada, eriti suuremael valimimahtudel.




Joonis 4: Testide võrdlus normaaljaotuse testimisel, kui andmed pärinevad kahemõõtmelisest jaotusest, mille üks komponent on normaaljaotusega, kuid teisele on rakendatud valemis (6) toodud funktsiooni, parameetriga $a = 1{,}2$

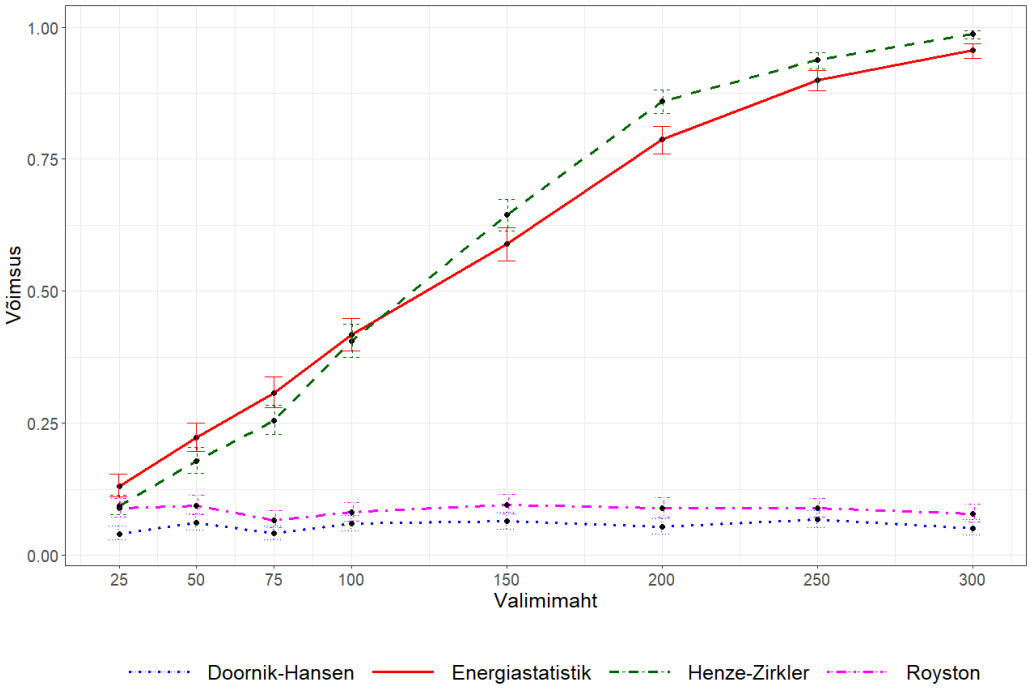
### 4.2.2 Kaheksamõõtmeline andmestik

Et uurida, kas mitmemõõtmelise normaaljaotuse testid töötavad ka sellistel andmetel, kus marginaaljaotused on normaaljaotused, kuid ühisjaotus ei ole mitmemõõtmeline normaaljaotus, sooritati testid kaheksamõõtmelisel andmestikul, mille üks komponent oli teise komponendi korrutis juhusliku suurusega, mis sai väärtuseks kas -1 või 1. Seega genereeriti iga valimimahu $n$ korral andmed järgmise programmiga:

```
X <- mvrnorm(n, mu = rep(0,8), diag(8))
S <- sample(c(-1, 1), size = n, replace = TRUE)
X[,2] <- X[,1]*S
valim <- X
```

Simulatsioonid viidi läbi väiksematel valimimahtudel kui kahemõõtmelisel juhul,

$n \in \{25; 50; 75; 100; 150; 200; 250; 300\}$. Samaks jäi igal valimimahul genereeritud valimite arv 1000. Tulemused on näha joonisel 5, kus on näha ka võimuste usaldusintervallid. Andmete genereerimise kood on välja toodud lisas 5




Joonis 5: Testide võrdlus normaaljaotuse testimisel, kui andmed pärinevad kaheksamõõtmelisest jaotusest, mille marginaaljaotused on normaaljaotused, kuid ühisjaotus ei ole mitmemõõtmeline normaaljaotus.

Simulatsioonidest selgus, et testid jagunesid selgelt kaheks - Doornik-Hanseni ja Roystoni testid ei suutnud normaaljaotusest kõrvalekallet tuvastada ning nende testide võimsus jäi kõigil eeltoodud valimimahtudel alla 0,1. Seevastu energiatest ning Henze-Zirkeli test suutsid tuvastada, et andmed ei ole mitmemõõtmelisest normaaljaotusest. Väiksemate valimimahtude korral saavutas natuke suurema võimsuse energiatest, kuid alates valimimahust 100, jäi see pisut Henze-Zirkeli testile alla. Mõlemad testid saavutasid testitavate valimimahtue juures võimsuse üle 0,8, energiatest saavutas selle valimimahul 250 ning Henze-Zirkeli test valimi suuruse 200 korral.

Seega saame öelda, et juhul, kus mitmemõõtmelise jaotuse kõik marginaaljaotused

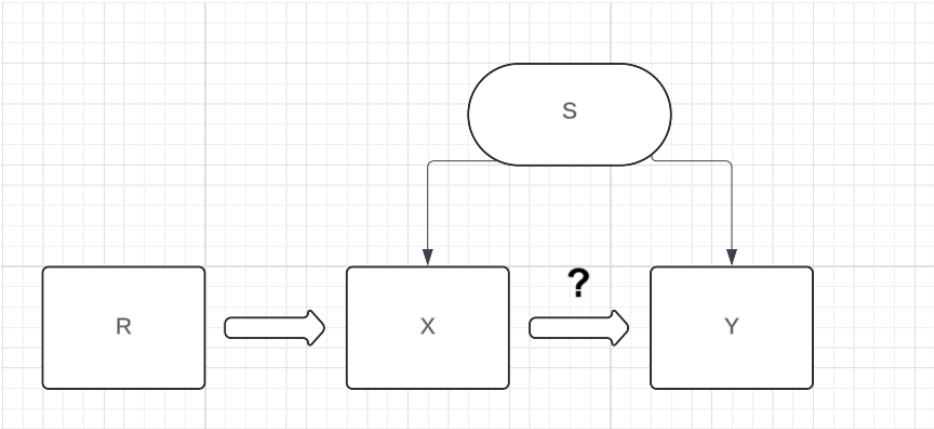
on normaaljaotused, kuid ühisjaotuseks ei ole mitmemõõtmeline normaaljaotus, on energiatest hea test erinevuse tuvastamiseks. Samuti tuleneb eelnevast analüüsist, et olenevalt andmete normaaljaotusest kõrvalekalde olemusest, peaks kasutama erinevaid teste, et saavutada parimad võimsused. Praktikas seda eriti aga teha ei saa, kuna ei teata andmete tegelikku jaotust. Sellisel juhul saab soovitada energiatesti kasutamist, sest see on kõikidel eeltoodud juhtudel hea võimalus kontrollida, kas andmed võiksid tuleneda (mitmemõõtmelisest) normaaljaotusest.

# 5 Energiatest struktuurivõrrandites

Struktuurivõrranditeks nimetatakse mudeleid, mille eesmärgiks on kirjeldada mingi tunnuse ning seda mõjutavate tegurite vahelisi põhjuslikke seoseid. Nende tööpõhimõte keskendub vaatluste asemel hoopis tunnuste kovariatsioonimustrite erinevuste minimiseerimisele. Enamik struktuurivõrrandeid eeldab, et tunnuste ühisjoatus on mitmemõõtmeline normaaljaotus. Kovariatsioonide minimiseerimiseks kasutatakse üldiselt suurima tõepära meetodit, kus mitmemõõtmelise normaaljaotuse eeldus oluliseks saabki. Struktuurivõrrandite nullhüpoteesiks on, et vaadeldud ning oletatava kovariatsioonimaatriksi vahel ei ole muid erinevusi peale erinevuse, mis on tingitud juhuslikkusest mitmemõõtmelisest normaaljaotusest valimi koostamisel. (Shipley, 2002, lk 100-115)

Vaatame struktuurivõrrandeid kujul, kus tunnus $R$ mõjutab tunnust $X$ ning tunnus $X$ omakorda tunnust $Y$, kuid tunnuseid $X$ ja $Y$ võib mõlemat mõjutada segav tunnus $S$. Soovime teada $X$ mõju tunnusele $Y$. Sellise struktuurivõrrandi kuju on toodud joonisel 6. Kui andmed ei ole mitmemõõtmelisest normaaljaotusest, võib tekkida probleem mudeli õigsuse ja interpreteeritavusega. Loome simulatsioonide jaoks eeltoodud kujul struktuurivõrrandi, kus tunnusel $X$ tegelikult puudub otsene mõju tunnusele $Y$. Samuti genereerime tunnuse $X$ selliselt, et tegemist ei oleks mitmemõõtmelise normaaljaotusega. Viime läbi simulatsioonid selliselt genereeritud andmetel valimimahuga $n = 50$ ning iteratsioonide arvuga 1000 (vt lisa 6). Selgub, et mudel tuvastab tunnuse $X$ nullist erineva mõju tunnusele $Y$ rohkematel juhtudel, kui klassikaliselt I liiki viga lubatakse teha. Täpsemalt tuvastati nullist erinev mõju 179 korral 1000 ehk I liiki viga tehti simuleeritud andmestiku korral 5% asemel 17,9% juhtudest.

Vaatame, kas energiatesti kasutamine struktuurivõrrandite puhul aitab tulemuste õigsust tagada. Selleks genereerime andmeid taas samal kujul nagu on näidatud joonisel 6, kuid enne iga mudeli hindamist kontrollime energiatestiga, kas andmed pärinevad normaaljaotusest. Kui energiatest ei tuvasta normaaljaotusest kõrvale-




Joonis 6: Vaadeldavate struktuurivõrrandite kuju, soovime uurida tunnuse $X$ mõju tunnusele $Y$

kallet, hindame mudeli ning loendame ainult selliste mudelite korral I liiki vigade arvu. Osutub, et sellisel viisil energiatesti kasutamine vähendab tõepoolest I liiki vea tegemise tõenäosust. Simuleeritud 1000 andmestikust jäeti kõrvale 167, nende puhul tuvastas energiatest kõrvalekalde mitmemõõtmelisest normaaljaotusest. Allesjäänud 833 andmestiku pealt hinnati struktuurivõrrandi mudelid. Nendest andmetest selgus, et mudel eksis vaid 31 korral ehk I liiki vea tegemise tõenäosuse hinnang langes 3,7 protsendini, olles 95% kindlusega vahemikus 2,5 kuni 5,2.

Järelikult saame öelda, et energiatest on hea viis vältida vigu struktuurivõrrandite mudelite hindamisel, kui kontrollida eelnevalt selle abil andmete vastavust mitmemõõtmelisele normaaljaotusele. Isegi kui energiatest ei tuvasta täielikult kõrvalekaldeid normaaljaotusest, aitab see vähendada mudeli tehtavate vigade arvu, muutes seeläbi analüüsi tulemusi paremaks ja usaldusväärsemaks.

# Kokkuvõte

Käesoleva töö eesmärgiks oli vaadelda energiastatistiku kasutusvõimalusi ning selgitada välja, kas ja millistes olukordades aitaks selle kasutamine statistilise analüüsi tulemusi paremaks muuta. Töös võrreldi energiatesti teiste tuntud testidega nii ühe- kui ka mitmemõõtmelise normaaljaotuse testimisel erinevate alternatiivsete jaotuste korral. Samuti vaadeldi selle kasutamist struktuurivõrrandite koostamisel, testides andmete pärinemist normaaljaotusest.

Ühemõõtmelisel juhul võrreldi energiatesti normaaljaotuse testimisel Kolmogorov-Smirnovi, Shapiro-Wilki ning Anderson-Darlingi testidega, kus andmete tegelikuks jaotuseks oli kas $t$-jaotus või normaaljaotuste segujaotus. Mõlema eeltoodud alternatiivi korral oli energiatesti võimsus võrreldav enamike teiste testidega, kuid olenevalt andmete tegelikust päritolust, osutus mõni teine test paremaks. Seega saab öelda, et ühemõõtmelisel juhul on energiatest piisavalt hea alternatiiv teistele testidele, kuid olenevalt andmete tegelikust jaotusest, võib energiatesti kasutamisel minna vaja suuremat valimimahtu.

Energiatesti võrreldi teiste testidega ka kahe- ja kaheksamõõtmelise andmestiku korral. Mitmemõõtmelise normaaljaotuse testimisel valiti võrdlemiseks Henze-Zirkeli, Roystoni ja Doornik-Hanseni testid. Kahemõõtmelises andmestikus muudeti ühte komponenti, et see ei vastaks normaaljaotusele. Sellisel meetodil saadud andmete peal normaaljaotust testides ei saavutanud energiatest küll parimaid tulemusi, kuid testi ei saanud pidada ka alternatiividest halvemaks. Kaheksamõõtmelist andmestikku modifitseeriti selliselt, et andmete marginaaljaotused olid normaaljaotused, kuid ühisjaotuseks ei olnud mitmemõõtmeline normaaljaotus. Tulemustest selgub, et sellistel andmetel töötab energiatest hästi ning tuvastab kõrvalekalded mitmemõõtmelisest normaaljaotusest.

Uurides energiastatistiku kasutusvõimalusi struktuurivõrrandite koostamisel ja hindamisel, selgus, et ka seal tasub energiatesti kasutamine ennast ära. Energiatesti sai kasutada, et kontrollida andmete pärinemist mitmemõõtmelisest normaajao-

tustest enne struktuurivõrrandi mudeli hindamist. Selliselt andmete päritolu kontrollimine aitas selgelt vähendada ekslike järelduste tegemist mudeli interpreteerimisel.

Seega saame energiatesti pidada heaks vahendiks nii ühe- kui ka mitmemõõtmelise normaaljaotuse testimisel. Kuigi mõni teine test võis saavutada soovitud võimsuse ka väiksemal valimimahul kui energiatest, ei jäänud see ka kõrvalekallete tuvastamisel hätta. Suureks energiatesti eeliseks on selle võimekus testida nii ühe- kui ka mitmemõõtmelisse normaaljaotusesse kuulumist.

# Kasutatud allikad


Korkmaz, Selcuk, Dincer Goksuluk ja Gokmen Zararsiz (2021). *MVN: Multivariate Normality Tests*. R package version 5.9. URL: https://CRAN.R-project.org/package=MVN.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rizzo, Maria L. ja Gabor J. Szekely (2022). *energy: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-11. URL: https://CRAN.R-project.org/package=energy.

Shipley, Bill (2002). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. 1. väljaanne. Cambridge University Press. ISBN: 9780521529211; 0521529212; 0511017723; 9780511017728; 0521791537; 9780521791533.

Székely, Gábor J. ja Maria L. Rizzo (2005). „A new test for multivariate normality". *Journal of Multivariate Analysis* 93, lk. 58–80. URL: https://www.sciencedirect.com/science/article/pii/S0047259X03002124?ref=pdf_download&fr=RR-2&rr=85b22da3adcd5422.

Székely, Gábor J. ja Maria L. Rizzo (2013). „Energy statistics: A class of statistics based on distances". *Journal of Statistical Planning and Inference* 143, lk. 1249–1272. URL: https://pages.stat.wisc.edu/~wahba/stat860public/pdf4/Energy/JSPI5102.pdf.

Székely, Gábor J. ja Maria L. Rizzo (2017). „The Energy of Data". *Annual Review of Statistics and Its Application* 4.1, lk. 447–479. DOI: 10.1146/annurev-statistics-060116-054026. URL: https://doi.org/10.1146/annurev-statistics-060116-054026.

Székely, Gábor J. ja Maria L. Rizzo (2023). *The Energy of Data and Distance Correlation*. 1. väljaanne. Taylor & Francis Group, LLC.

Tabak, J. (2014). *Geometry: The Language of Space and Form*. Facts on File math library. Facts On File, Incorporated. ISBN: 9780816068760. URL: https://books.google.ee/books?id=r0HuPiexnYwC.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

# Lisa 1. Monte Carlo meetodil Bernoulli jaotuse testimine

```
n <- 3
R <- 1000
stat <- function(valim){
  liige1 <- 2/3 * sum(valim*0.9 + abs(valim-1)*0.1)
  liige2 <- 2*0.1*0.9
  liige3 <- 1/9 * (sum(abs(valim[1]-valim)) +
                   sum(abs(valim[2]-valim)) +
                   sum(abs(valim[3]-valim)))
  return(3*(liige1-liige2-liige3))
}
valim <- c(0,0,1)
h0 <- stat(valim)
suuremaid <- 0
statistikud <- c()
for (i in 1:R) {
  x_tärn <- rbinom(n,1,0.1)
  s <- stat(x_tärn)
  if (s >= h0) {suuremaid = suuremaid+1}
  statistikud <- append(statistikud, s)
}
q <- quantile(statistikud, 0.95)
p <- suuremaid/R
print(h0) # statisik
print(p) # p-väärtus
print(h0<q) # kas soovitud jaotus
```

# Lisa 2. Normaaljaotuse testimine, kui andmed pärinevad t-jaotusest

```
library(energy)
library(nortest)
# testide võimsuste vektorid
v_e = v_ks = v_sw = v_ad = rep(NA,7)
# testide usalduspiiride vektorid
min_e = max_e = min_ks = max_ks = min_sw = max_sw = min_ad = max_ad = rep(NA,7)
n <- c(100,500,1000,2000, 3000, 4000, 5000) # valimimahud
m <- 1000 # iteratsioonide arv
alpha <- 0.05
for (i in 1:7){
  # testide alternatiivse hüpoteesi vastu võtmiste arvud
  H_ks =  H_e = H_sw = H_ad = 0
  for(j in 1:m){
    valim <- rt(n[i],15) # valim t-jaotusest vabadusastmete arvuga 15
    if(ks.test(valim, "pnorm")$p.value <= alpha){
      H_ks <- H_ks+1
    }
    if(normal.test(valim, R = 100)$p.value <= alpha){
      H_e <- H_e+1
    }
    if(shapiro.test(valim)$p.value <= alpha){
      H_sw <- H_sw+1
    }
    if(ad.test(valim)$p.value <= alpha){
      H_ad <- H_ad +1
    }
```

```
  }
  v_e[i] <- H_e/m
  v_ks[i] <- H_ks/m
  v_sw[i] <- H_sw/m
  v_ad[i] <- H_ad/m
  u_e <- binom.test(H_e, m, conf.level = 1-alpha)$conf.int
  min_e[i] <- u_e[1]
  max_e[i] <- u_e[2]
  u_ks <- binom.test(H_ks, m, conf.level = 1-alpha)$conf.int
  min_ks[i] <- u_ks[1]
  max_ks[i] <- u_ks[2]
  u_sw <- binom.test(H_sw, m, conf.level = 1-alpha)$conf.int
  min_sw[i] <- u_sw[1]
  max_sw[i] <- u_sw[2]
  u_ad <- binom.test(H_ad, m, conf.level = 1-alpha)$conf.int
  min_ad[i] <- u_ad[1]
  max_ad[i] <- u_ad[2]
}
# lõpilik andmestik
andmed <- data.frame(min = min_e, voimsus = v_e, max = max_e,
        maht = n, test = "Energiastatistik")
andmed <- rbind(andmed, data.frame(min = min_ks,voimsus = v_ks,
        max = max_ks, maht = n, test= "Kolmogorov-Smirnov"))
andmed <- rbind(andmed, data.frame(min = min_sw, voimsus = v_sw,
        max = max_sw, maht = n, test = "Shapiro-Wilk"))
andmed <- rbind(andmed, data.frame(min = min_ad, voimsus = v_ad,
        max = max_ad, maht = n, test = "Anderson-Darling"))
```

# Lisa 3. Normaaljaotuse testimine juhul, kui andmed pärinevad normaaljaotuste segujaotusest

```
library(energy)
library(nortest)
# testide võimsuste vektorid
v_e = v_ks = v_sw = v_ad = c()
# testide usalduspiiride vektorid
min_e = max_e = min_ks = max_ks = min_sw = max_sw = min_ad = max_ad = c()
n <- c(100, 1000)
valimi_keskväärtused <- c(1, 0.8, 0.6, 0.4, 0.2, 0)
m <- 1000
alpha <- 0.05
for (z in valimi_keskväärtused){
  for (i in n){
    H_ks = H_e = H_sw = H_ad = 0
    for(j in 1:m){
      valim <- rbinom(i,1,prob = 0.5)
      valim[valim == 0] <- rnorm(sum(valim == 0), z, 1)
      valim[valim == 1] <- rnorm(sum(valim == 1), -z, 1)
      if(ks.test(valim, "pnorm", 0, 1)$p.value <= alpha){
        H_ks <- H_ks+1
      }
      if(normal.test(valim, R = 100)$p.value <= alpha){
        H_e <- H_e+1
      }
      if(shapiro.test(valim)$p.value <= alpha){
        H_sw <- H_sw+1
      }
```

```
      if(ad.test(valim)$p.value <= alpha){
        H_ad <- H_ad +1
      }
    }
    v_e <- append(v_e,(H_e/m))
    v_ks <- append(v_ks,(H_ks/m))
    v_sw <- append(v_sw,(H_sw/m))
    v_ad <- append(v_ad,(H_ad/m))
  }
}
split_ks <- split(v_ks, rep(1:6, each = length(v_ks) / 6))
df_ks <- as.data.frame(do.call(cbind, split_ks))
df_ks <- cbind(df_ks, n)
df_ks["test"] <- "Kolmogorov-Smirnov"
split_en <- split(v_e, rep(1:6, each = length(v_e) / 6))
df_en <- as.data.frame(do.call(cbind, split_en))
df_en <- cbind(df_en, n)
df_en["test"] <-"Energiatest"
split_ad <- split(v_ad, rep(1:6, each = length(v_ad) / 6))
df_ad <- as.data.frame(do.call(cbind, split_ad))
df_ad <- cbind(df_ad, n)
df_ad["test"] <- "Anderson-Darling"
split_sw <- split(v_sw, rep(1:6, each = length(v_sw) / 6))
df_sw <- as.data.frame(do.call(cbind, split_sw))
df_sw <- cbind(df_sw, n)
df_sw["test"] <- "Shapiro-Wilk"
# lõplik andmestik
tabel <- dplyr::bind_rows(df_ks, df_ad, df_en, df_sw)
tabel2 <- reshape2::melt(tabel, id = c("n","test"))
```

# Lisa 4. Normaaljaotuse testimine kahemõõtmelisel mitte-normaaljaotusest andmestikul

```
library(energy)
library(MVN)
exponent <- function(a, pow) (abs(a)^pow)*sign(a) #funktsioon andmete tekitamiseks
n <- c(200, 300, 400, 500, 600, 700, 800, 900, 1000)
m <- 1000
alpha <- 0.05
andmed2 = data.frame(row.names = c("aste", "min", "voimsus", "max", "maht", "test"))
astmed <- c(1.1, 1.2)
for (a in astmed){
  # võimsuste vektorid
  v_e = v_hz = v_r = v_dh = rep(NA, 9)
  # usalduspiiride vektorid
  min_e = max_e = min_hz = max_hz = min_r = min_dh = max_r = max_dh = rep(NA,9)
  for (i in 1:9){
    H_hz = H_e = H_r = H_dh = 0
    for(j in 1:m){
      v1 <- rnorm(n[i], -1, 1)
      v2 <- exponent(rnorm(n[i], 1, 1),a)
      valim <- cbind(v1,v2)
      if (mvnorm.test(valim, R=100)$p.value <= alpha){
        H_e <- H_e +1
      }
      if((mvn(valim, mvnTest = 'hz')$multivariateNormality)$p <= alpha){
        H_hz <- H_hz+1
      }
      if((mvn(valim, mvnTest = 'royston')$multivariateNormality)$p <= alpha){
```

```
      H_r <- H_r+1
    }
    if((mvn(valim, mvnTest = 'dh')$multivariateNormality)$p <= alpha){
      H_dh <- H_dh+1
    }
  }
  v_e[i] <- H_e/m
  v_hz[i] <- H_hz/m
  v_r[i] <- H_r/m
  v_dh[i] <- H_dh/m
  u_e <- binom.test(H_e, m, conf.level = 1-alpha)$conf.int
  min_e[i] <- u_e[1]
  max_e[i] <- u_e[2]
  u_hz <- binom.test(H_hz, m, conf.level = 1-alpha)$conf.int
  min_hz[i] <- u_hz[1]
  max_hz[i] <- u_hz[2]
  u_r <- binom.test(H_r, m, conf.level = 1-alpha)$conf.int
  min_r[i] <- u_r[1]
  max_r[i] <- u_r[2]
  u_dh <- binom.test(H_dh, m, conf.level = 1-alpha)$conf.int
  min_dh[i] <- u_dh[1]
  max_dh[i] <- u_dh[2]
}
# lõplikud andmed
andmed2 <- rbind(andmed2, data.frame(aste = a, min = min_e,
      voimsus = v_e, max = max_e, maht = n, test = "Energiastatistik"))
andmed2 <- rbind(andmed2, data.frame(aste = a, min = min_hz,
      voimsus = v_hz, max = max_hz, maht = n, test= "Henze-Zirkler"))
andmed2 <- rbind(andmed2, data.frame(aste = a, min = min_r,
```

```
        voimsus = v_r, max = max_r, maht = n, test = "Royston"))
  andmed2 <- rbind(andmed2, data.frame(aste = a, min = min_dh,
        voimsus = v_dh, max = max_dh, maht = n, test = "Doornik-Hansen"))
}
```

# Lisa 5. Normaaljaotuse testimine kaheksamõõtmelisel andmestikul

```
library(energy)
library(MVN)
library(MASS)
# võimsuste vektorid
v_e = v_hz = v_r = v_dh = rep(NA, 8)
# usalduspiiride vektorid
min_e = max_e = min_hz = max_hz = min_r = min_dh = max_r = max_dh = rep(NA,8)
n <- c(25, 50, 75, 100, 150, 200, 250, 300)
m <- 1000
alpha <- 0.05
for (i in 1:8){
  H_hz = H_e = H_r = H_dh = 0
  for(j in 1:m){
    X <- mvrnorm(n[i], mu = rep(0,8), diag(8))
    S <- sample(c(-1, 1), size = n[i], replace = TRUE)
    X[,2] <- X[,1]*S
    valim <- X
    if (mvnorm.test(valim, R=100)$p.value <= alpha){
      H_e <- H_e +1
    }
```

```
    if((mvn(valim, mvnTest = 'hz')$multivariateNormality)$p <= alpha){
      H_hz <- H_hz+1
    }
    if((mvn(valim, mvnTest = 'royston')$multivariateNormality)$p <= alpha){
      H_r <- H_r+1
    }
    if((mvn(valim, mvnTest = 'dh')$multivariateNormality)$p <= alpha){
      H_dh <- H_dh+1
    }
  }
  v_e[i] <- H_e/m
  v_hz[i] <- H_hz/m
  v_r[i] <- H_r/m
  v_dh[i] <- H_dh/m
  u_e <- binom.test(H_e, m, conf.level = 1-alpha)$conf.int
  min_e[i] <- u_e[1]
  max_e[i] <- u_e[2]
  u_hz <- binom.test(H_hz, m, conf.level = 1-alpha)$conf.int
  min_hz[i] <- u_hz[1]
  max_hz[i] <- u_hz[2]
  u_r <- binom.test(H_r, m, conf.level = 1-alpha)$conf.int
  min_r[i] <- u_r[1]
  max_r[i] <- u_r[2]
  u_dh <- binom.test(H_dh, m, conf.level = 1-alpha)$conf.int
  min_dh[i] <- u_dh[1]
  max_dh[i] <- u_dh[2]
}
# lõplik andmestik
andmed8 <- data.frame(row.names = c("min", "voimsus", "max", "maht", "test"))
```

```
andmed8 <- rbind(andmed8, data.frame(min = min_e, voimsus = v_e, max = max_e,
        maht = n, test = "Energiastatistik"))
andmed8 <- rbind(andmed8, data.frame(min = min_hz,voimsus = v_hz, max = max_hz,
        maht = n, test= "Henze-Zirkler"))
andmed8 <- rbind(andmed8, data.frame(min = min_r, voimsus = v_r, max = max_r,
        maht = n, test = "Royston"))
andmed8 <- rbind(andmed8, data.frame(min = min_dh, voimsus = v_dh, max = max_dh,
        maht = n, test = "Doornik-Hansen"))
```

# Lisa 6. Energiatesti kasutamine struktuurivõrrandites

```
library(lavaan)
set.seed(1111)
exponent <- function(a, pow) (abs(a)^pow)*sign(a)
mudel="y~x
x~r
x~~y
"
n = 50
k1 = 0 # vigade arv algselt
# ilma normaaljaotust testimata
for (i in 1:1000){
  s=rnorm(n)
  r=rnorm(n)
  x=+5*s-20*r+c(exponent(rnorm(n,-3,1),1.25),exponent(rnorm(n,3,1),1.25))
  y=0*x-8*s+rnorm(n,0,1)
  andmed=data.frame(x,y,r)
```

```
  m=lavaan(mudel, andmed, auto.var=TRUE)
  if (summary(m)$pe[1,8]<0.05) {
    k1=k1+1
  }
}
n = 50
k2 = 0 # vigade arv esmalt normaaljaotust testides
alles = 1000
# testides iga andmestiku puhul, kas see on mitmemõõtmelisest normaaljaotusest
for (i in 1:1000){
  s=rnorm(n)
  r=rnorm(n)
  x=+5*s-20*r+c(exponent(rnorm(n,-3,1),1.25),exponent(rnorm(n,3,1),1.25))
  y=0*x-8*s+rnorm(n,0,1)
  andmed=data.frame(x,y,r)
  p_vaartus = energy::mvnorm.test(andmed, R=100)$p
  if (p_vaartus > 0.05){
    alles = alles - 1
    m=lavaan(mudel, andmed, auto.var=TRUE)
    if (summary(m)$pe[1,8]<0.05) {
      k2=k2+1
    }
  }
}
```

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Karmeli Kaasik,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Energiakauguse rakendamine normaaljaotuse eelduse testimisel", mille juhendaja on Märt Möls, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karmeli Kaasik

14.05.2024