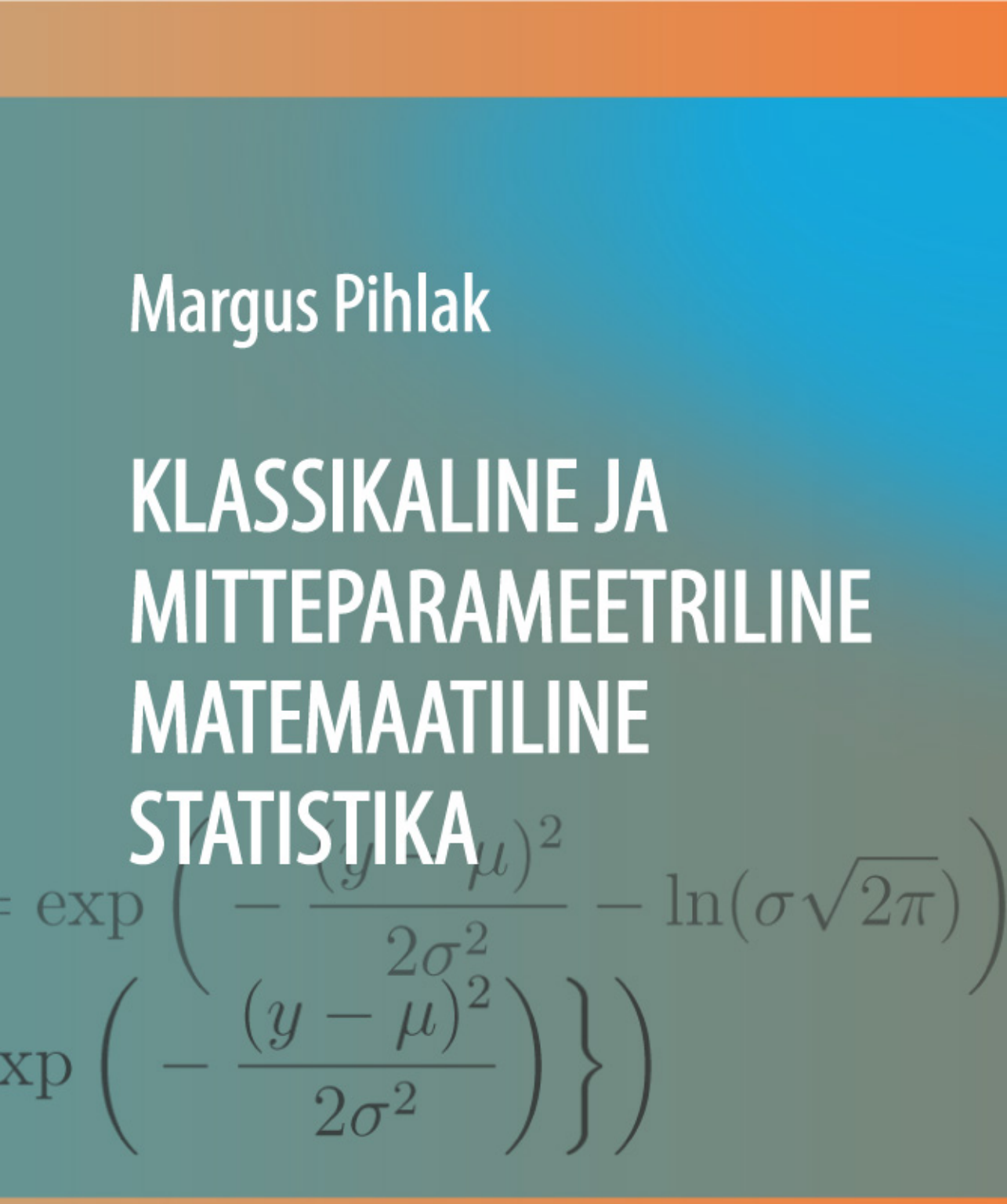
Margus Pihlak

KLASSIKALINE JA MITTEPARAMEETRILINE MATEMAATILINE STATISTIKA



Margus Pihlak

KLASSIKALINE JA MITTEPARAMEETRILINE MATEMAATILINE STATISTIKA

Õpik kõrgkoolidele

TTÜ KIRJASTUS

Tallinn 2018

Ilmunud riikliku programmi
„Eestikeelsete kõrgkooliõpikute koostamine ja väljaandmine 2013–2017“
toetusel

Käesoleva õpiku väljaandmist toetasid

Haridus- ja Teadusministeerium

ARCHIMEDES

Retsenseerinud
Tartu Observatooriumi vanemteadur Elmo Tempel
Tartu Observatooriumi juhtivteadur Enn Saar

Kaane kujundanud Tiia Eikholm

ISBN 978-9949-83-349-8 (köites)
ISBN 978-9949-83-299-6 (pdf)

Sisukord

Sisukord **3**

Eessõna **7**

1. Klassikaline statistika **10**

1.1. Sissejuhatus statistilisse analüüsi 10

1.1.1. Matemaatilise statistika olemus 10

1.1.2. Klassikalise statistika eeldused 12

1.2. Statistikud ja hinnangud 14

1.2.1. Punkthinnang 15

1.2.2. Hinnang momentide meetodil 19

1.2.3. Suurima tõepära (STP) hinnang 22

1.2.4. Vahemikhinnang 26

1.3. Statistiliste hüpoteeside kontroll 32

1.3.1. Hüpoteeside kontrolli ideoloogiline olemus 33

1.3.2. Statistilise testi võimsus 38

1.3.3. Näiteid erinevatest statistilistest hüpoteesidest . . 41

1.3.4. Statistiliste hüpoteeside kontrollimine tarkvara R abil 65

1.4. Ülesanded 68

2. Andmeanalüüs **75**

2.1. Objekt-tunnus maatriks 75

2.2. Üldise lineaarse mudeli struktuur 77

2.2.1. Mudeli parameetrite leidmine vähimruutude meetodil 78

2.2.2. Üldise lineaarse mudeli diagnostika 82

2.2.3. Mudeli jääkide analüüs 90

2.3. Dispersioonanalüüsi alused 93

2.3.1. Ühefaktoriline dispersioonanalüüs 94

2.3.2. Kahefaktoriline dispersioonanalüüs 97

2.3.3. Dispersioonanalüüs tarkvarade MS Excel ja R abil 100

2.4. Üldistatud lineaarsed mudelid 103

2.4.1. Eksponentsiaalsete jaotuste pere 103

2.4.2. Seosefunktsioonid 107

2.4.3. Logistilised mudelid 108

2.4.4. Üldistatud lineaarsete mudelite diagnoosimine . . 112

2.4.5. Üldistatud lineaarsete mudelite koostamine ning diagnoosimine tarkvara R abil 119

2.5. Faktoranalüüs 122

2.5.1. Tunnuste jagamine faktoriteks 123

2.5.2. Peakomponentide meetod 128

2.6. Ülesanded 133

3. Mitteparameetriline statistika 140

3.1. Mitteparameetrilise statistika erinevused võrreldes klassikalise statistikaga 140

3.2. Ekstremaalsete väärtuste teooria 141

3.2.1. Järkstatistikute jaotused 142

3.2.2. Ekstremaalsete väärtuste jaotused 145

3.3. Tunnustevahelised astakkorrelatsioonid 147

3.3.1. Spearmanni kordaja 148

3.3.2. Kendalli kordaja 153

3.4. Mitteparameetrilised statistilised testid 155

3.4.1. Wilcoxoni astakute test 155

3.4.2. Kvantiili test 161

3.4.3. Kruskal-Wallise H-test 162

3.4.4. Friedmanni test 165

3.4.5. Kolmogorov-Smirnovi test 167

3.4.6. Anderson-Darlingi test 173

3.4.7. Wilk-Šapiro test 174

3.5. Taasvaliku meetodid 176

3.5.1. *Jackknife*-meetod 176

3.5.2. *Bootstrap*-meetod 179

3.5.3. Vahemikhinnangute leidmine *bootstrap*-meetodil 182

3.5.4. *Bootstrap*-hinnangud tarkvara R abil 187

3.5.5. Permutatsiooni test 189

3.6. Ülesanded 193

4. Juhuslik vektor ja Bayesi statistika **200**

4.1. Mitmemõõtmelise statistika alused 201

4.1.1. Diskreetne juhuslik vektor 201

4.1.2. Pidev juhuslik vektor 212

4.1.3. Tinglik keskväärtus ja regressioonanalüüs 215

4.2. Bayesi meetodid 219

4.2.1. Täistõenäosus ja Bayesi valem klassikalisel kujul 219

4.2.2. Tinglikustamise võte 221

4.2.3. Bayesi valem pideval juhul 222

4.2.4. Statistiliste hüpoteeside kontrollimine Bayesi meetodiga 228

4.3. Markovi ahelad statistikas 241

4.3.1. Bayesi teoreem 241

4.3.2. Ülevaade Markovi ahelatest 246

4.3.3. Näiteid MCMC-mudelitest 252

4.3.4. Tarkvara OpenBUGS 261

4.4. Ülesanded 270

Lisad **276**

Ülesannete vastused ja lahendused **280**

Kirjandus **284**

Aineregister **288**

Eessõna

Suurem osa loodus- ja täppisteadusest on seotud mõõtmistega. Mõõtmised aga tekitavad andmeid ning need omakorda suure hulga informatsiooni. Mida sellega peale hakata? Mida sellest informatsioonist järeldada? Nendele küsimustele annab paljuski vastuse matemaatiline statistika. See on teadus, mille eesmärk on genereerida andmetest jaotusi. Ehk teisiti öelduna, matemaatilise statistika sisendiks on mõõtmistulemused ning väljundiks jaotus. Selle jaotuse põhjal leitakse tõenäosus, et mingi näitaja tegelik väärtus kuulub teatud vahemikku.

Nõudlus matemaatilise statistika järele on Eesti ülikoolides viimastel aastatel kasvanud. Enamikel erialadel nii Tallinna Tehnikaülikoolis, Tartu Ülikoolis kui ka Eesti Maaülikoolis on tõenäosusteooria ja matemaatilise statistika põhikursus kohustuslik, kuid sobivat eestikeelset õpperaamatut pole seni olnud. Selle õpiku üks eesmärke on täita see lünk. Teine eesmärk on olla teatud määral teejuht neile, kellel on vaja rakendada matemaatilist statistikat oma uurimistöödes, seda eelkõige loodus- ja inseneriteadustes.

Õpik koosneb neljast peatükist. Esimene peatükk on mõeldud üheks õppevahendiks matemaatilise statistika põhialuste omandamisel. See käsitleb klassikalist statistikat ehk nn Fisheri statistikat. Peatüki esimeses osas esitatakse klassiklise statistika eeldused. Neist põhiline on parameetrilisuse eeldus, kus eeldatakse, et uuritav suurus on teadaoleva jaotusega. Eesmärk on hinnata selle jaotuse parameetreid. Teises osas vaadeldakse jaotuse parameetrite erinevaid hinnanguid: punkt-, vahemik-, suurima tõepära hinnangut. Selle peatüki kolmas osa on pühendatud

statistiliste hüpoteeside kontrollile. Õpiku teises peatükis tutvustatakse andmemassiividega töötamist. Selles kirjeldatakse andmeanalüüsi põhimeetodeid. Samuti tutvustatakse, kuidas analüüsida andmestikku tarkvarade MS Excel ning R abil.

Kolmas ja neljas peatükk sobivad õppevahendiks ülikooli doktoriõppes loetavatele kursustele, täpsemalt mitteparameetrilise statistika ja Bayesi statistika omandamisel. Kolmas peatükk on pühendatud mitteparameetrilisele ehk jaotuste vabale statistikale. Need on juhud, kus klassikalise statistika eeldused ei kehti. Näiteks kui tuleb teha statistiline otsustus väikese valimi põhjal. Esmalt võrreldakse klassikalist statistikat mitteparameetrilisega. Teine osa selles peatükis on pühendatud järkstatistikutele ja ekstremaalsetele väärtustele. Kolmandas osas vaadeldakse juhtusid, millal lineaarne korrelatsioonikordaja ei anna adekvaatset infot. Selle peatüki neljandas osas kirjeldatakse mitmeid mitteparameetrilisi teste: Wilcoxoni test, Friedmanni test jms. Viiendas osas antakse ülevaade taasvaliku meetoditest, millest peamised on *bootstrap*- ja *jackknife*-meetodid.

Neljas peatükk käsitleb Bayesi statistikat, mida võib vaadelda tõenäosusteooria ja matemaatilise statistika vahelise sümbioosina. Esimeses osas antakse ülevaade mitmemõõtmelisest statistikast. Käsitletakse selliseid mõisteid nagu ühisjaotus, tinglik jaotus, tinglik keskväärtus ja regressioonikordaja. Teises osas esitatakse Bayesi statistika põhiolemus. See seisneb selles, et teoreetiliste eelteadmiste põhjal eeldame, et jaotuse parameetril on mingi kindel jaotus. Küsimus on, kui palju muudab seda eeldust empiiriline andmestik. Teisisõnu, kuivõrd erineb parameetri jaotus enne ja pärast eksperimenti? Selle peatüki kolmas osa on pühendatud meetodile nimega MCMC (ingl ***M**arkov **C**hain **M**onte **C**arlo*), milles on ühendatud teooria ja empiirika. Meetod põhineb Bayesi statistikal. Antakse ülevaade Markovi ahelatest ning tuuakse näiteid erinevatest MCMC-mudelitest.

Raamatu igas peatükis on rõhutatud tõenäosusteooria ja matemaatilise statistika terviklikkust. Kõikide peatükkide lõpus on sellele osale vastavad harjutusülesanded. Neid on nii teoreetilise kui ka praktilise kallakuga.

Lugejalt eeldatakse, et ta on läbinud ülikoolis tõenäosusteoorias põhikursuse või omandanud õpiku [34] I osa materjali. Samuti tulevad kasuks maatriksalgebra ning matemaatilise analüüsi teadmised ja oskused, mille kohta on õpikud [35] ja [27].

Õpikus esinevad põhitähistused.

1) Arvestades mitme tarkvara (Matlab, Maple, R jms) süntaksiga, on arvu täisosa eraldajaks õpikus „ . “. Näiteks 1.025.

2) Tõenäosust tähistatakse traditsiooniliselt kui P, juhusliku suuruse keskväärtust ja dispersiooni tähistavad traditsioonilised tähistused, vastavalt E ja D.

3) Seda, et juhuslikku suurust X kirjeldab jaotusfunktsioon F, tähistatakse kui $X \sim F$.

4) Kombinatsioonide hulka n elemendist k elemendi kaupa tähistatakse kui C_n^k.

5) Arvu $e = 2.7182...$ astet näitab lühend $\exp$.

6) Tõestuse lõppu tähistab õpikus sümbol $\square$.

7) Maatriksi transponeerimist tähistab sümbol $^\intercal$.

8) Hulga A elementide hulka ehk hulga A võimsust tähistatakse kui $| A |$.

Siinkohal tänan Tartu Observatooriumi juhtivteadurit Enn Saart ja sama asutuse vanemteadurit Elmo Templit asjalike retsensioonide ning kasulike märkuste eest. Need aitasid oluliselt parandada õpiku sisu. Samuti tänan Tallinna Tehnikaülikooli lektorit Mati Väljast, kes aitas õpikuga seotud tehniliste probleemide lahendamisel. Veel avaldan tänu Tartu Ülikooli tehnoloogiainstituudi vanemteadurile Ülo Maivälile, kes andis kasulikke näpunäiteid Bayesi meetodite rakenduste osas.

Head aine omandamist!

Tallinnas maikuus 2018

1. peatükk
Klassikaline statistika

Matemaatilise statistika võib jagada tinglikult kaheks: klassikaliseks ja mitteparameetriliseks statistikaks. Ehk jaotustepõhiseks ning n-ö jaotuste vabaks statistikaks. Selles peatükis käsitleme klassikalist statistikat. Esmalt aga tutvume statistilise analüüsi põhiolemusega.

1.1. Sissejuhatus statistilisse analüüsi

1.1.1. Matemaatilise statistika olemus

Igal teadusel on oma baasmõisted, neist saab vastav teadus alguse. Matemaatilises statistikas on baasmõisteteks üldkogum ja valim.

Definitsioon 1.1. Hulka, mille kohta tehakse otsustus, nimetatakse üldkogumiks.

Definitsioon 1.2. Üldkogumi alamhulka, mille põhjal tehakse otsustus, nimetatakse valimiks.

Eesmärk on uurida suure hulga ühetüübiliste objektide (ehk üldkogumi) mingit iseloomulikku tunnust. Näiteks:

1) meid huvitava detaili läbimõõtu,

2) Peipsi järve rääbiste kehakaalu,

3) mingil maanteelõigul aset leidnud liiklusõnnetuste arvu kuus.

Hindamaks meid huvitavaid näitajaid tehakse ühetaoliste objektide hulgast valim. Tähistagem üldkogumit kui U ja valimit kui u. Definitsiooni 1.2 põhjal $u \subseteq U$ ehk valim on üldkogumi alamhulk.

Definitsioon 1.3. Elementide hulka valimis u nimetatakse valimi mahuks.

Olgu kogu käesoleva õpiku jooksul valimi mahu tähistuseks n. Käesoleva õpiku tähistustes $\mid u \mid = n$.

Statistikat võib nimetatada **matemaatiliseks**, kui $u \subset U$. Seega teeb matemaatiline statistika suurema hulga kohta otsustuse väiksema hulga põhjal. Selle otsustuse tegemisel kasutatakse tõenäosusteooria tulemusi. Tõenäosusteooria ja matemaatiline statistika moodustavad ühe terviku. See tähendab, et ühe sisend on teise väljund. Kui tõenäosusteooria sisendiks on teoreetilised jaotused ning väljundiks empiirilised andmed, siis matemaatilise statistika eesmärk on genereerida andmetest jaotusi. Seda genereerimist tehakse valimi põhjal.

Matemaatilise statistika tegemise ehk statistilise analüüsi võib jagada 4 etappi.

1) Esimene etapp seisneb katse planeerimises. Selle käigus tuleb formuleerida probleemid, millele tahetakse saada seletust.

2) Teine etapp seisneb andmete kogumises. Selle etapi jooksul tuleb määratleda, mis on antud probleemile vastav üldkogum ning koostada võimalikult adekvaatne valim.

3) Käesolev õpik on pühendatud kolmandale etapile. Selle käigus viiakse läbi statistilise analüüsi see osa, mis puudutab matemaatikat (täpsemalt tõenäosusteooriat). Leitakse valimi põhjal saadud hinnangud ning tehakse nende abil prognoose üldkogumi kohta.

4) Neljas etapp puudutab analüüsi tulemuste interpreteerimist. Nende tulemuste põhjal antakse seletus esimeses etapis formuleeritud probleemidele loodus- ja inseneriteadustes, sotsioloogias ning muudes valdkondades.

Asume järgnevalt lähemalt uurima erinevaid meetodeid statistilise analüü-

si läbiviimisel. Enne seda aga tuleb sõnastada klassikalise statistika eeldused.

1.1.2. Klassikalise statistika eeldused

Klassikaline statistika on matemaatilise statistika haru, mis on leidnud enim rakendust. Selle põhialused rajas 20. sajandi esimesel poolel Briti statistik ja bioloog Roland Aylmer Fisher (1890–1962). Vaatamata teaduse tormilisele arengule pole klassikaline statistika oma aktuaalsust kaotanud tänapäevalgi. Klassikalises statistikas peegeldub suuresti statistilise analüüsi põhiolemus: valimi põhjal hinnangute koostamine ja nende alusel tegelikkust puudutavate hüpoteeside püstitamine. Formuleerime alljärgnevalt klassikalise matemaatilise statistika eeldused. Need võib jagada kolme gruppi.

1° Üldkogum U on lõpmatu ja valim u on selle lõplik alamhulk, see tähendab valimi maht $|u| = n < \infty$.

2° Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$. Eeldame, et selle valimi elemendid on sõltumatud, see tähendab valik on tagasipanekuga. Igal elemendil olgu valimisse kaasamise tõenäosus $\dfrac{1}{n}$.

3° Parameetrilisuse eeldus, kus eeldame, et $X_i \sim F(\Theta)$, $i = 1, 2, ..., n$. Jaotusfunktsioon F kirjeldab üldkogumit U. Vektorit

$$\Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}$$

nimetatakse selle jaotusfunktsiooni parameetrite vektoriks. Eesmärk on hinnata parameetreid θ_j, $j = 1, 2, ..., k$.

Esitame vektori Θ struktuuri mõningate jaotuste puhul.

Näide 1.1. Olgu X_i normaaljaotusega, mida tähistame kui $X_i \sim \mathcal{N}(\mu, \sigma)$. Siis tihedusfunktsioon

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right), \quad \mu > 0.$$

Seega on jaotuse parameetrite vektoriks

$$\Theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}.$$

Näide 1.2. Olgu X_i binoomjaotusega, mida tähistame kui $X_i \sim B(n,p)$. Selle jaotuse puhul tõenäosusfunktsioon

$$P(X_i = k) = C_n^k p^k (1-p)^{n-k}, \quad p \in (0;1).$$

Seega on jaotuse parameetrite vektoriks

$$\Theta = \begin{pmatrix} n \\ p \end{pmatrix}.$$

Näide 1.3. Olgu X_i Poissoni jaotusega, mida tähistame kui $X_i \sim Po(\lambda)$. Selle jaotuse puhul tõenäosusfunktsioon

$$P(X_i = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad \lambda > 0.$$

Seega jaotuse parameetrite vektor $\Theta = \lambda$.

Näide 1.4. Olgu juhuslik suurus X_i eksponentjaotusega, mida tähistame kui $X_i \sim \mathcal{E}(\nu)$. Siis tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x < 0, \\ \nu \exp(-\nu x), \text{ kui } x \geq 0 \end{cases}, \quad \nu > 0.$$

Selle jaotuse parameetrite vektor $\Theta = \nu$.

Näide 1.5. Olgu X_i geomeetrilise jaotusega, mida tähistame kui $X_i \sim Geo(p)$. Sellisel juhul saame tõenäosusfunktsiooniks

$$P(X_i = k) = p(1-p)^{k-1}, \quad p \in (0;1)$$

ning jaotuse parameetrite vektoriks $\Theta = p$.

1.2. Statistikud ja hinnangud

Järgnevalt uurime lähemalt, kuidas hinnata valimi põhjal üldkogumit kirjeldava jaotuse parameetreid.

Käsitleme valimi u elemente kui juhuslikke suurusi. Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$, kus $X_i \sim F$, $i = 1, 2, ..., n$. Esindagu seda valimit juhuslik suurus X, mis on valimi $\mathbf{X}$ suvalise elemendiga identse jaotusega koopia. Edaspidi hakkame kirjeldama konkreetset valimit tema esindaja kaudu. Olgu $\mathbf{x} = (x_1, x_2, ..., x_n)^\intercal$ selle juhusliku valimi mingi realisatsioon. Realistatsiooni all mõtleme konkreetseid mõõtmistulemusi. Meie eesmärk on hinnata valimi $\mathbf{X}$ abil üldkogumit kirjeldava jaotuse F parameetreid. Selleks toome sisse mõiste statistik.

Definitsioon 1.4. Statistikuks nimetatakse valimi põhjal moodustatud juhuslikku suurust hindamaks jaotuse parameetrit θ_j, $j = 1, 2, ..., k$.

Olgu parameetri θ_j hinnang $\widehat{\theta_j}$. Vastavat statistikut tähistame edaspidi kui $\widehat{\theta_j} = T(\mathbf{X})$. Selle statistiku realisatsiooniks $T(\mathbf{x})$ on mingi ratsionaalarv, vahemik, lõik või poollõik, mis sõltub mõõtmistulemustest. Meie edaspidiseks põhiülesandeks saab olema realisatsioonile $T(\mathbf{x})$ vastava tõenäosuse α leidmine. Kui statistiku $T(\mathbf{X})$ jaotus on diskreetne, siis

$$\alpha = P(T(\mathbf{X}) = \theta_j).$$

See tähendab, et meid huvitab, millise tõenäosusega α võtab meid huvitav parameeter väärtuse θ_j. Kui statistiku $T(\mathbf{X})$ jaotus on pidev, siis huvitab meid tõenäosus

$$\alpha = P(T(\mathbf{X}) \in \mathcal{H}_\alpha),$$

kus $\mathcal{H}_\alpha$ tähistab parameetri θ_j väärtuste lõiku, poollõiku või vahemikku. Edaspidi kasutame nende tõenäosuste tähistamiseks enamasti lühemat kirjapilti

$$\alpha = P(\widehat{\theta_j} = \theta_j) \text{ või } \alpha = P(\widehat{\theta_j} \in \mathcal{H}_\alpha).$$

Tõenäosuse α ning piirkonna $\mathcal{H}_\alpha$ leidmine on matemaatilise statistika üks peamisi ülesandeid.

1.2.1. Punkthinnang

Punkthinnang on teistest hinnangutest lihtsaim. See annab hinnatava parameetri tõenäoseima asukoha reaalteljel. Punkthinnangu headuse uurimisel rakendatakse juhusliku suuruse keskväärtuse ja dispersiooni omadusi. Neid saab meelde tuletada näiteks õpikust [34] (lk 60–68).

Punkthinnangut võib nimetada peaaegu ideaalseks, kui tal on kaks head omadust. Esimene neist on hinnangu nihketus.

Definitsioon 1.5. Punkthinnangut $\widehat{\theta}_j = T(\mathbf{X})$ nimetatakse nihketa hinnanguks (ingl *unbiased estimator*), kui

$$E(T(\mathbf{X})) = \theta_j$$

sõltumata valimi mahust n.

Teine hea omadus puudutab hinnangu mõjusust ehk täpsustuvust.

Definitsioon 1.6. Punkthinnangut $\widehat{\theta}_j$ nimetatakse mõjusaks (ingl *consistent*) ehk täpsustuvaks, kui

$$\lim_{n\to\infty} D(\widehat{\theta}_j) = 0.$$

Mõjusust saab defineerida ka nõrgema tingimuse ehk tõenäosuse järgi koondumise abil.

Definitsioon 1.7. Punkthinnangut $\widehat{\theta}_j$ nimetatakse mõjusaks ehk täpsustuvaks, kui $\forall \epsilon > 0$ korral

$$\lim_{n\to\infty} P(\mid \widehat{\theta}_j - \theta_j \mid < \epsilon) = 1.$$

Definitsiooni 1.6 mõjususe tingimusest järeldub definitsiooni 1.7 tingimus. Vastupidine järelduvus üldjuhul ei kehti.

Lahti seletatult tähendab hinnangu mõjusus seda, et valimi mahu kasvades hinnangu hajuvus kahaneb. Hajuvuse ehk dispersiooniga mõõdetakse hinnangu efektiivsust. Olgu meil parameetri θ_j hinnangud $\widehat{\theta}_j^1$ ning $\widehat{\theta}_j^2$.

Definitsioon 1.8. Kui hinnangud $\widehat{\theta}_j^1$ ning $\widehat{\theta}_j^2$ on nihketa ning

$$D(\widehat{\theta}_j^1) < D(\widehat{\theta}_j^2),$$

siis öeldakse, et hinnang $\widehat{\theta}_j^1$ on efektiivsem kui hinnang $\widehat{\theta}_j^2$.

Näide 1.6. Olgu meil valimi $\mathbf{X} = (X_1, X_2, X_3)^\intercal$ põhjal saadud hinnangud

$$\widehat{\theta}_j^1 = \frac{X_1 + X_2 + X_3}{3} \text{ ning } \widehat{\theta}_j^2 = \frac{X_1 + X_2}{2}.$$

Võrdleme nende hinnangute efektiivsusi. Klassikalise statistika eelduste 2° ja 3° põhjal saame, et

$$D(\widehat{\theta}_j^1) = D\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{9}D(X_1+X_2+X_3) = \frac{1}{9}3D(X_1) = \frac{D(X_1)}{3}$$

ning

$$D(\widehat{\theta}_j^2) = D\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}D(X_1 + X_2) = \frac{1}{4}2D(X_1) = \frac{D(X_1)}{2}.$$

Seega $D(\widehat{\theta}_j^1) < D(\widehat{\theta}_j^2)$ ning hinnang $\widehat{\theta}_j^1$ on efektiivsem kui hinnang $\widehat{\theta}_j^2$.

Hinnangu viga jaguneb kaheks: juhuveaks ja süstemaatiliseks veaks. Juhuviga on tingitud mõõtmise ebatäpsusest, süstemaatiline viga aga valest mõõtmise metoodikast.

Definitsioon 1.9. Suurust

$$b = E(\widehat{\theta}_j) - \theta_j$$

nimetatakse süstemaatiliseks veaks ehk hinnangu nihkeks.

Definitsioon 1.10. Suurust

$$s_{viga} = \frac{\sigma_{X_i}}{\sqrt{n}},$$

kus σ_{X_i} tähistab juhusliku suuruse X_i standardhälvet, nimetatakse X_i keskväärtuse hinnangu standardveaks.

Süstemaatilist viga iseloomustab nihe, juhuviga aga standardhälve. Defineerime järgnevalt punkthinnangu keskmise ruutvea parameetri θ_j hindamisel.

Definitsioon 1.11. Punkthinnangu $\widehat{\theta}_j$ keskmine ruutviga

$$\mathrm{MSE}(\widehat{\theta}_j) = E(\widehat{\theta}_j - \theta_j)^2.$$

Lühend MSE tuleb inglise keelest (***M**ean **S**quare **E**rror*). Me saame avaldada keskmise ruutvea järgmiselt:

$$\mathrm{MSE}(\widehat{\theta}_j) = E(\widehat{\theta}_j - E(\widehat{\theta}_j) + E(\widehat{\theta}_j) - \theta_j)^2 =$$

$$= E\{(\widehat{\theta}_j - E(\widehat{\theta}_j))^2 + 2(\widehat{\theta}_j - E(\widehat{\theta}_j))(E(\widehat{\theta}_j) - \theta_j) + (E(\widehat{\theta}_j) - \theta_j)^2\}$$

$$= E\{\widehat{\theta}_j - E(\widehat{\theta}_j)\}^2 + \{\theta_j - E(\widehat{\theta}_j)\}^2 = D(\widehat{\theta}_j) + b^2,$$

sest

$$2E\{(\widehat{\theta}_j - E(\widehat{\theta}_j))(E(\widehat{\theta}_j) - \theta_j)\} = 0.$$

Seega koosneb suurus MSE

1) juhuvea komponendist $D(\widehat{\theta}_j)$,

2) süstemaatilise vea komponendist b^2.

Toome ühe klassikalise näite punkthinnangu nihketusest ja mõjususest. Olgu meil valim, kus $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Meie eesmärk on anda statistikuga $T(\mathbf{X})$ punkthinnangud parameetritele μ ja σ.

1) Olgu meil keskväärtuse μ punkthinnanguks

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.1}$$

Veendume, et hinnang (1.1) on nihketa. Keskväärtuse omadusi rakendades saame, et

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu.$$

Saime, et $E(\overline{x}) = \mu$, millest järeldub hinnangu (1.1) nihketus.

Veendume, et hinnang (1.1) on ka mõjus. Dispersiooni omadusi rakendades ning valimisse kaasamise sõltumatuse eeldusi kasutades (millise võrdusmärgi juures?) saame, et

$$D(\overline{x}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Saime, et $D(\overline{x}) = \dfrac{\sigma^2}{n}$. Seega keskmine ruutviga

$$MSE(\overline{x}) = \frac{\sigma^2}{n}.$$

Piirväärtusest

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

järeldub hinnangu (1.1) mõjusus.

Arvestades normaaljaotuse invariantsust lineaarteisenduse suhtes, saab eelpool leitust teha järgmise järelduse.

Järeldus 1.1. Olgu juhuslik suurus $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$, siis juhuslik suurus $\overline{x} \sim \mathcal{N}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

2) Vaatame dispersiooni σ^2 punkthinnangut

$$\overline{s^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2. \tag{1.2}$$

Tegemist on ju ruutkauguse aritmeetilise keskmisega. Kuid hinnang (1.2) on nihkega. Selle statistiku keskväärtus

$$E(\overline{s^2}) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu + \mu - \overline{x})^2 =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E((X_i-\mu)^2 - 2E(\overline{x}-\mu)(X_i-\mu) + E(\overline{x}-\mu)^2 \right) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i-\mu)^2 - 2E(\overline{x}-\mu)\sum_{i=1}^{n}(X_i-\mu) + nE(\overline{x}-\mu)^2 \right) =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i-\mu)^2 - nE(\overline{x}-\mu)^2 \right).$$

Dispersiooni definitsioonist järeldub, et $E(X_i-\mu)^2 = \sigma^2$. Järeldusest 1.1 saame, et $E(\overline{x}-\mu)^2 = \dfrac{\sigma^2}{n}$. Seega

$$E(\overline{s^2}) = \frac{1}{n}(n\sigma^2 - \sigma^2) = \sigma^2 - \frac{\sigma^2}{n},$$

millest järeldubki, et hinnang (1.2) on nihkega. Selle nihe $b = -\dfrac{\sigma^2}{n}$. Kuigi varsti veendume, et tegemist on teatavas mõttes hea hinnanguga, hindab statistik (1.2) tegelikku dispersiooni alla. Tekib küsimus: mida teha, et nihe kõrvaldada? Osutub, et selleks tuleb hinnangu (1.2) nimetajat ühe võrra vähendada.

Lause 1.1. Olgu meil $X_i \sim \mathcal{N}(\mu,\sigma)$, $i = 1, 2, ..., n$. Siis statistik

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{x})^2 \tag{1.3}$$

on nihketa punkthinnanguks dispersioonile σ^2.

Jätame hinnangu (1.3) nihketuse lugejale tõestada ülesandes 1.3. Arvestades hinnangu s^2 nihketust, saame, et standardvea nihketa hinnang

$$s_{viga} = \frac{s}{\sqrt{n}}.$$

1.2.2. Hinnang momentide meetodil

Momentide meetod üldistab punkthinnangut. Tuletame meelde momentide tõenäosusteooriast tuntud momentide mõistet. Olgu meil juhuslik

suurus $X \sim F$, millele vastab pideval juhul tihedusfunktsioon $f(x)$ ning diskreetsel juhul tõenäosusfunktsioon $p_i = P(X = x_i)$, $i = 1, 2, ..., n$. Siis selle juhusliku suuruse j-ndat järku moment

$$m_j = E(X^j) = \int_{-\infty}^{\infty} x^j f(x) dx$$

pideval juhul. Kui juhusliku suuruse X kõikvõimalike väärtuste hulk on lõplik, siis

$$m_j = \sum_{i=1}^{n} x_i^j p_i.$$

Kui aga loenduv, siis

$$m_j = \sum_{i=1}^{\infty} x_i^j p_i.$$

Jaotuse parameetrite hindamisel ei saa kasutada momentide meetodit, kui vastav integraal või rida ei koondu.

Momentides peitub juhusliku suuruse X kohta palju infot. Nii saame leida 1. momendi abil selle juhusliku suuruse keskväärtuse, 2. momendi abil tema dispersiooni, 3. momendi abil aga asümmeetria kordaja (ingl *skewness*). Juhusliku suuruse momendid aga sõltuvad jaotuse F parameetritest $\theta_1, \theta_2, ..., \theta_k$. Seega võib j-ndat järku momenti käsitleda kui funktsiooni, mis sõltub k muutujast

$$m_j = \mu_j(\theta_1, \theta_2, ..., \theta_k).$$

Valimi $(X_1, X_2, ..., X_n)^\intercal$ põhjal leitakse j-ndat järku momendi nihketa hinnangud

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} X_i^j,$$

mille abil koostatakse võrrandisüsteem

$$\mu_j(\theta_1, \theta_2, ..., \theta_k) = \overline{x_j}, \quad j = 1, 2, ..., k. \tag{1.4}$$

Definitsioon 1.12. Võrrandisüsteemi (1.4) lahendit $(\theta_1^*, \theta_2^*, ..., \theta_k^*)^\intercal$ nimetatakse parameetrite vektori $\Theta = (\theta_1, \theta_2, ..., \theta_k)^\intercal$ hinnanguks momentide meetodil.

Rakendame momentide meetodi hinnangut erinevatele jaotustele.

Näide 1.7. Olgu meil valim, mida esindab normaaljaotusele $\mathcal{N}(\mu, \sigma)$ alluv juhuslik suurus X. Siis

$$m_1 = \mu \text{ ning } m_2 = \mu^2 + \sigma^2.$$

Parameetrite μ ja σ hinnangud momentide meetodil on seega võrrandisüsteemi

$$\begin{cases} \mu = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \\ \mu^2 + \sigma^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \end{cases}$$

lahendid.

Olgu meil mõõdetud näiteks 8 mehe pikkused (cm):

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
176	191	182	184	180	179	175	192

Eeldame, et meeste pikkus allub normaaljaotusele. Parameetri μ hinnang

$$\mu^* = \overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i \approx 182.$$

Standardhälbe parameetri hinnang

$$\sigma^* \approx \sqrt{\frac{1}{8}\sum_{i=1}^{8} x_i^2 - 182^2} \approx 5.9.$$

Näide 1.8. Olgu meil valim, mida esindab lõigul $[a; b]$ ühtlasele jaotusele alluv juhuslik suurus X. Siis selle juhusliku suuruse tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x \notin [a; b] \\ \dfrac{1}{b-a}, \text{ kui } x \in [a; b]. \end{cases}$$

Leidmaks parameetrite a ja b hinnangud momentide meetodil tuleb lahendada võrrandisüsteem

$$\begin{cases} \dfrac{a+b}{2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \\ \dfrac{a^2+ab+b^2}{3} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2. \end{cases}$$

Ühtlasele jaotusele vastavate momentide m_1 ja m_2 leidmine jäägu lugejate ülesandeks. Süsteemi lahendid on järgmised:

$$a = m_1 \pm \sqrt{3(m_2 - m_1^2)} \text{ ning } b = m_1 \mp \sqrt{3(m_2 - m_1^2)}.$$

Näide 1.9. Olgu meil valim, mida esindav juhuslik suurus allub eksponentjaotusele $\mathcal{E}(\nu)$. Siis on parameetri ν hinnang momentide meetodil leitav võrrandiga

$$\frac{1}{\nu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Taas jäägu lugeja veenduda, et eksponentjaotuse puhul 1. järku moment

$$m_1 = \int_0^\infty x\nu \exp(-\nu)dx = \frac{1}{\nu}.$$

1.2.3. Suurima tõepära (STP) hinnang

Hinnangu nime ingliskeelne lühend on MLE (ingl ***M**aximum **L**ikelihood **E**stimation*). Tegemist on hinnanguga, mis sisaldab kogu infot meid huvitavast juhuslikust suurusest. Seda infot kirjeldab pideval juhul tihedusfunktsioon ning diskreetsel juhul tõenäosusfunktsioon. Defineerime nende funktsioonide abil tõepära funktsiooni, mida tähistame kui L. Teeme seda eraldi pideval ja diskreetsel juhul. Pideval juhul on juhusliku suuruse X võimalike väärtuste hulk lõik, poollõik või vahemik. Diskreetsel juhul aga on see väärtuste hulk ülimalt loenduv.

Pideval juhul avaldub tõepära funktsioon L kui

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} f(\Theta, x_i) \tag{1.5}$$

ning diskreetsel juhul kui

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} P(X = x_i). \tag{1.6}$$

Eesmärk on leida tõepära funktsiooni $L(\Theta, \mathbf{x})$ maksimumkohad. Maksimumkohtade leidmine taandub enamasti funktsiooni tuletise nullkohtade leidmisele, nagu on teada matemaatilisest analüüsist. Antud juhul tuleb lahendada järgmine võrrandisüsteem:

$$\begin{cases} \dfrac{\partial}{\partial\theta_1} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) = 0 \\ \dfrac{\partial}{\partial\theta_2} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) = 0 \\ \vdots \\ \dfrac{\partial}{\partial\theta_k} L(\theta_1, \theta_2, ..., \theta_k, \mathbf{x}) = 0. \end{cases} \tag{1.7}$$

Selle süsteemi lahendiks on meid huvitava jaotuse parameetrite STP hinnang $\Theta^* = (\theta_1^*, \theta_2^*, ..., \theta_k^*)^\intercal$ ehk parameetrite sellised väärtused, mis teevad valimi realisatsiooni $\mathbf{x}$ korral tõepära funktsiooni L väärtuse maksimaalseks.

Maksimeerimiseks tuleb leida korrutise tuletis, nagu võib tähele panna tõepärafunktsiooni definitsioonist. Funktsioonide korrutise diferentseerimine on tehniliselt väga raske ülesanne. Korrutise saab aga teha summaks logaritmimise abil. Summa tuletis on teatavasti aditiivne. Järelikult muutub logaritmimisega ülesanne tehniliselt märksa kergemaks ülesande olemust muutmata. Leiame logaritmitud tõepära funktsiooni pideval juhul

$$l(\Theta, \mathbf{x}) = \ln(L(\Theta, \mathbf{x})) = \ln\left(\prod_{i=1}^{n} f(\Theta, x_i)\right) = \sum_{i=1}^{n} \ln(f(\Theta, x_i)).$$

Analoogiliselt saab logaritmida diskreetsel juhul.

Sageli on võrrandite süsteemi (1.7) lahendamine väga keeruline ülesanne. Võib juhtuda, et sellel puudub analüütiline lahend. Rakendame STP hinnangut jaotustele, mille puhul on süsteemi lahendamine suhteliselt hõlbus.

Näide 1.10. Olgu meil valim, kus $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, ..., n$. Tõenäosusteooriast on teada, et $E(X_i) = \mu$ ja $D(X_i) = \sigma^2$. Normaaljaotusele vastav tõepärafunktsioon

$$L(\Theta, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Logaritmides saame, et

$$l(\Theta, \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Leiame STP hinnangu parameetrile μ. Selleks leiame osatuletise

$$\frac{\partial l(\mu, \sigma^2, \mathbf{x})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Võrdsustades antud osatuletise suurusega 0, saame, et

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0,$$

millest järeldub, et

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

Saime, et normaaljaotuse parameetri μ STP hinnanguks on aritmeetiline keskmine $\overline{x}$. Leiame osatuletise parameetri σ^2 järgi asendades parameetri μ tema STP hinnanguga. Saame, et

$$\frac{\partial l(\overline{x}, \sigma^2, \mathbf{x})}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2} = 0.$$

Pärast mõningaid teisendusi saame, et

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

Seega on parameetri σ^2 STP hinnanguks s_*^2. See hinnang aga on teatavasti nihkega. Normaaljaotuse puhul saime parameetrite vektori STP hinnanguks $\Theta^* = (\overline{x}, s_*^2)^\intercal$.

Näide 1.11. Olgu meil valim, kus X_i, $i = 1, 2, ..., n$, allub Poissoni jaotusele parameetriga $\lambda > 0$. Tõenäosusteooriast on teada, et selle jaotuse korral $E(X_i) = D(X_i) = \lambda$. Poissoni jaotusele vastav logaritmitud tõepära funktsioon

$$l(\lambda, \mathbf{x}) = \ln(\lambda) \sum_{i=1}^{n} x_i - \ln(x_i!) - n\lambda.$$

Võrdsustades osatuletise parameetri λ järgi suurusega 0, saame, et

$$\frac{\partial l(\lambda, \mathbf{x})}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0,$$

millest järeldub, et

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

Seega on parameetri λ STP hinnanguks aritmeetiline keskmine $\overline{x}$.

Näide 1.12. Olgu meil valim, kus X_i, $i = 1, 2, ..., n$, allub eksponentjaotusele parameetriga $\nu > 0$. Sellisel juhul võrduvad nii keskväärtus kui ka standardhälve suurusega $\frac{1}{\nu}$. Logaritmitud tõepära funktsioon

$$l(\nu, \mathbf{x}) = n \ln(\nu) - \nu \sum_{i=1}^{n} x_i.$$

Võrdsustades osatuletise parameetri ν järgi suurusega 0, saame, et

$$\frac{\partial l(\nu, \mathbf{x})}{\partial \nu} = \frac{n}{\nu} - \sum_{i=1}^{n} x_i = 0,$$

millest järeldub, et

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

Seega on parameetri ν STP hinnanguks aritmeetilise keskmise $\overline{x}$ pöördväärtus.

Näide 1.13. Olgu meil valim, kus X_i, $i = 1, 2, ..., n$, allub Rayleigh' jaotusele parameetriga $h > 0$. Siis tõenäosuse tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x < 0, \\ 2h^2 x \exp(-h^2 x^2), \text{ kui } x \geq 0. \end{cases}$$

Rayleigh' jaotus leiab laialdast rakendust loodusteaduses ja tehnikas, näiteks tomograafias. Ka elanikkonna tulude jaotus on sageli lähedane Reighleigh' jaotusele. Sellele vastav keskväärtus ja dispersioon avalduvad järgmiselt:

$$E(X) = \frac{\sqrt{\pi}}{2h} \text{ ja } D(X) = \frac{4-\pi}{4h^2}.$$

Logaritmitud tõepära funktsioon

$$l(h, \mathbf{x}) = n\ln(2) + 2n\ln(h) + \sum_{i=1}^{n} \ln(x_i) - h^2 \sum_{i=1}^{n} x_i^2.$$

Võrdsustades selle funktsiooni osatuletise parameetri h järgi suurusega 0, saame, et

$$\frac{\partial l(h, \mathbf{x})}{\partial h} = \frac{2n}{h} - 2h \sum_{i=1}^{n} x_i^2 = 0,$$

millest järeldub, et

$$h = \frac{1}{\sqrt{\overline{x^2}}},$$

kus

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

Seega on Reighleigh' jaotuse parameetri STP hinnanguks arimeetilise ruutkeskmise ruutjuure pöördväärtus.

1.2.4. Vahemikhinnang

Vahemikhinnang võtab arvesse nii statistiku keskväärtuse kui ka standardhälbe. Kõikidest hinnangutest kasutatakse praktikas enim just vahemikhinnangut. Meie eesmärk on leida α-usaldusintervall parameetrile θ_j. Defineerime esmalt selle intervalli.

Definitsioon 1.13. Lõiku I_α nimetatakse α-usaldusintervalliks parameetrile θ_j, kui

$$P(\theta_j \in I_\alpha) = \alpha$$

Teisisõnu, α-usaldusintervall katab parameetrit θ_j tõenäosusega α.

Suurust α nimetatakse definitsioonis 1.13 usaldusnivooks. Olgu $I_\alpha = [u, U]$. Siis nimetatakse suurust u alumiseks usalduspiiriks (ingl *lower bound*) ning suurust U ülemiseks usalduspiiriks (ingl *upper bound*). Definitsioonist 1.13 järeldub, et

$$P(\theta_j < u) = P(\theta_j > U) = \frac{1-\alpha}{2}.$$

Kui statistiku jaotusele vastava tihedusfunktsiooni graafik on sümmeetriline y-telje suhtes, siis $u = -U$.

Vahemikhinnangu puhul on oluline mõiste, mida nimetatakse juhusliku suuruse α-kvantiiliks.

Definitsioon 1.14. Juhusliku suuruse X väärtust x_α nimetatakse tema α-kvantiiliks, kui

$$P(X \leq x_\alpha) = \alpha.$$

Funktsiooni $F^{-1}(\alpha)$ nimetatakse juhusliku suuruse X kvantiili funktsiooniks. Seega

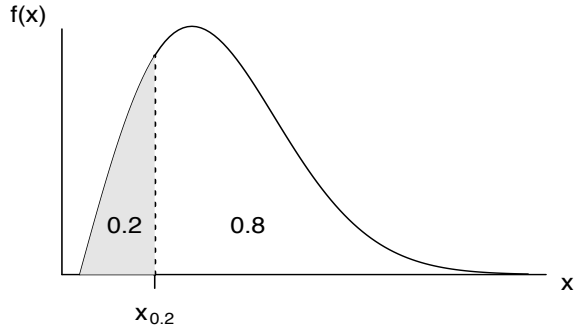
$$x_\alpha = F^{-1}(\alpha).$$

Definitsioonist 1.14 järeldub, et jaotusfunktsiooni pidevuse korral on kvantiili funktsioon jaotusfunktsiooni pöördfunktsioon. Analoogiliselt saame defineerida α-täiendkvantiili.

Definitsioon 1.15. Juhusliku suuruse X väärtust $\overline{x}_\alpha$ nimetatakse tema α-täiendkvantiiliks, kui

$$P(X > \overline{x}_\alpha) = \alpha.$$

Seega on juhusliku suuruse α-kvantiil tema $1-\alpha$-täiendkvantiiliks ehk $x_\alpha = \overline{x}_{1-\alpha}$. Joonis 1.1 illustreerib Rayleigh' jaotusega juhusliku suuruse kvantiili ja täiendkvantiili.



Joonis 1.1. Rayleigh' jaotusega ($h = 0.3$) juhusliku suuruse 0.2-kvantiil ja 0.8-täiendkvantiil

Vaatame vahemikhinnangut, mis põhineb normaalsel aproksimatsioonil ehk normaaljaotusega lähendamisel. Need lähendid võib jagada kaheks.

Esimene neist baseerub tsentraalse piirteoreemi rakendamisel. Siinkohal peab meelde tuletama rakenduste mõttes üliolulist tõenäosusteooria tulemust. Tsentraalse piirteoreemi väide tähendab mingi juhuslike suuruste jada koondumist jaotuse järgi.

Definitsioon 1.16. Öeldakse, et juhuslike suuruste jada $X_1 \sim F_1, X_2 \sim F_2, ..., X_n \sim F_n, ...$ koondub jaotuse järgi juhuslikuks suuruseks $X \sim F$, kui

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0$$

iga väärtuse x korral, millal jaotusfunktsioon $F(x)$ on pidev.

Teoreem 1.1. (tsentraalne piirteoreem) Olgu meil juhuslikud suurused $X_1, X_2, ..., X_n$ sõltumatud ja sama jaotusega (s.s.j). Vastav ingliskeelne lühend on *i.i.d.* (***I**ndependent **I**dentically **D**istributed*). Olgu

$$E(X_i) = \mu < \infty \text{ ja } D(X_i) = \sigma^2 < \infty.$$

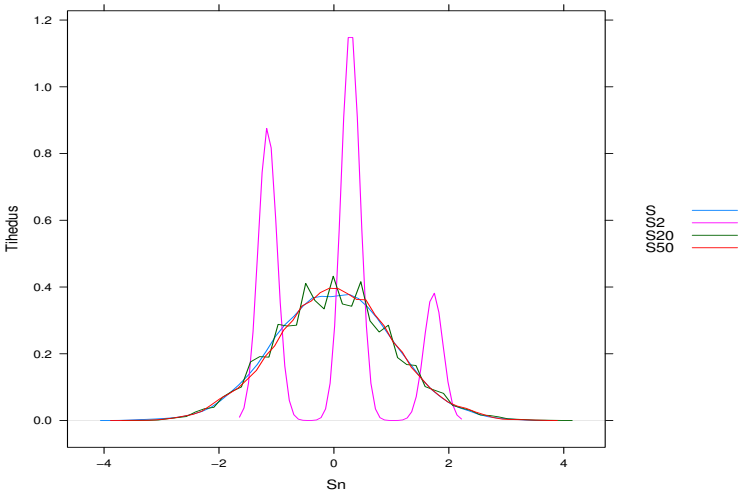
Siis juhuslike suuruste jada $\left\{ S_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \right\}$ koondub jaotuse järgi juhuslikuks suuruseks $S \sim \mathcal{N}(0,1)$.

Jagades juhusliku suuruse S_n avaldises lugeja ning nimetaja valimi mahuga n, saame lugejasse aritmeetilise keskmise $\overline{x}$ ning teoreetilise keskmise μ vahe. Seega piisavalt suure n (rusikareegli kohaselt $n \geq 30$) puhul juhuslik suurus

$$Z = \frac{\overline{x} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1),$$

kus juhuslikud suurused $X_1, X_2, ..., X_n$ on sõltumatud ja sama jaotusega (s.s.j) ning $\mu = E(X_i)$ ja $\sigma^2 = D(X_i)$.

Demonstreerime tsentraalse piirteoreemi toimimist binoomjaotuse baasil. Selleks genereeriti juhusliku arvude generaatoriga 1000 juhuslikku suurust $X_n \sim B(n, 0.4)$ ning $S \sim \mathcal{N}(0,1)$. Antud juhul saame teoreemi 1.1 tähistustes $\mu = n \cdot 0.4$ ning $\sigma = \sqrt{n \cdot 0.4 \cdot 0.6}$. Joonisel 1.2 on kujutatud juhusliku suuruse S_n tihedusfunktsioonide graafikuid $n = 2$, $n = 20$ ning $n = 50$ korral.



Joonis 1.2. Juhusliku suuruse S_n tihedusfunktsioonide võrdlus standardse normaaljaotuse tihedusfunktsiooniga

Tsentraalse piirteoreemi rakendamist binoomjaotusele võib sõnastada järgmise tulemusena.

Teoreem 1.2. (De Moivre-Laplace'i piirteoreem) Binoomjaotusega juhuslike suuruste jada $X_1 \sim B(1,p), X_2 \sim B(2,p), ..., X_n \sim B(n,p), ...$

koondub jaotuse järgi juhuslikuks suuruseks, mis allub normaaljaotusele

$$\mathcal{N}(np, \sqrt{np(1-p)}).$$

Teine lähend põhineb asjaolul, et juhusliku suuruse

$$Z = \frac{T(\mathbf{X}) - E\{T(\mathbf{X})\}}{\sqrt{D\{T(\mathbf{X})\}}}$$

jaotus läheneb standardsele normaaljaotusele, kui $n \to \infty$. Seda eeldusel, et juhuslikul suurusel X_i, $i = 1, 2, ..., n$, leiduvad keskväärtus ja dispersioon ning funktsioonil $T(\mathbf{X})$ leiduvad tuletised kuni järguni 3.

Leiame tsentraalset piirteoreemi rakendades α-usaldusintervalli keskväärtusele μ. Selleks lähtume tema nihketa ja mõjusast hinnangust, milleks on aritmeetiline keskmine $\overline{x}$. Eesmärk on leida selline suurus ϵ_α, mille korral

$$P(\mu \in [\overline{x} - \epsilon_\alpha; \overline{x} + \epsilon_\alpha]) = \alpha.$$

Saame, et

$$P(\overline{x} - \epsilon_\alpha \leq \mu \leq \overline{x} + \epsilon_\alpha) = \alpha \Leftrightarrow P(-\epsilon_\alpha \leq \overline{x} - \mu \leq \epsilon_\alpha) = \alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n} \leq \frac{\overline{x} - \mu}{\sigma}\sqrt{n} \leq \frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) = \alpha = F\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) - F\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right),$$

kus F tähistab juhusliku suuruse

$$Z = \frac{\overline{x} - \mu}{\sigma}\sqrt{n}$$

jaotusfunktsiooni. Tsentraalse piirteoreemi põhjal aga on juhuslik suurus Z ligikaudu standardse normaaljaotusega, s.t normaaljaotusega, mille keskväärtus on 0 ja standardhälve 1. Tõenäosusteooriast on teada, et normaaljaotusel puudub jaotusfunktsioon anaüütilisel kujul. Selle jaotuse puhul kasutatakse tõenäosuse leidmiseks Laplace'i veafunktsiooni

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right) dt.$$

Sellel funktsioonil on järgmised omadused:

1° kui $X \sim \mathcal{N}(0, 1)$, siis tema jaotusfunktsioon $F(x) = 0.5 + \Phi(x)$;

2° kui $X \sim \mathcal{N}(\mu, \sigma)$, siis $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$;

3° Laplace'i veafunktsioon on paaritu ehk $\Phi(-x) = -\Phi(x)$;

4° piirväärtus

$$\lim_{x\to\infty} \Phi(x) = 0.5.$$

Funktsiooni Φ omaduste 2° ja 3° ning tsentraalse piirteoreemi põhjal saame, et

$$F\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) - F\left(-\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) \approx 2\Phi\left(\frac{\epsilon_\alpha}{\sigma}\sqrt{n}\right) = \alpha.$$

Arvestades definitsiooni 1.14, saame, et

$$\epsilon_\alpha = \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}.$$

Kui $\alpha = 0.95$, siis $\Phi^{-1}\left(\frac{\alpha}{2}\right) \approx 1.96$ (vaata lisa 1) ning α-usaldusintervall

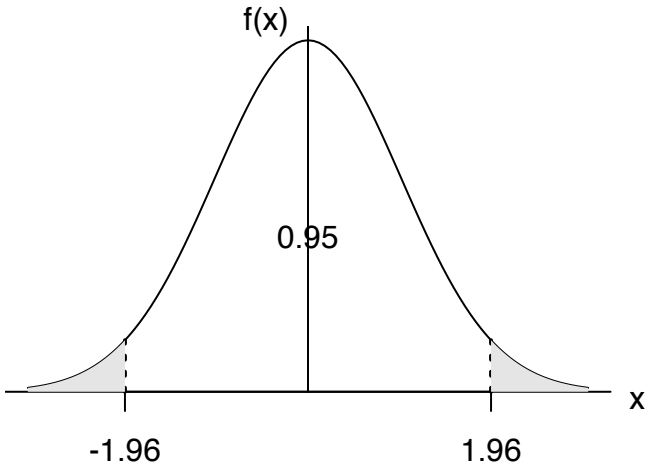
$$I_\alpha \approx \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Kui võtta $\alpha = 0.99$, siis $\Phi^{-1}\left(\frac{\alpha}{2}\right) \approx 2.58$ ning

$$I_\alpha \approx \left[\overline{x} - 2.58\frac{\sigma}{\sqrt{n}}; \overline{x} + 2.58\frac{\sigma}{\sqrt{n}}\right].$$

Mida suurem on valimi maht n, seda täpsem on α-usaldusintervalli hinnang. Kui suurendada tõenäosust α, siis suureneb usaldusintervalli pikkus.

Allpool olev standardse normaaljaotuse tihedusfunktsiooni graafik illustreerib, kuidas on saadud 0.95-usaldusintervalli korral arv 1.96.



Joonis 1.3. Suuruse $\Phi^{-1}\left(\frac{0.95}{2}\right)$ leidmine

Näide 1.14. Olgu meil valim $\mathbf{X}$. Me teeme 9 mõõtmist, mille põhjal saame sellele valimile realisatsiooni

$$\mathbf{x} = (1.1, 1.8, 1.9, 2.1, 2.2, 2.5, 1.3, 1.9, 1.8)^\intercal.$$

Olgu uuritava juhusliku suuruse standardhälve $\sigma = 0.5$. Valimi realisatsiooni põhjal saame aritmeetiliseks keskmiseks $\overline{x} \approx 1.8$. Tegeliku keskväärtuse 0.95-usaldusintervall

$$I_{0.95} \approx [1.5; 2.2]$$

ning 0.99-usaldusintervall

$$I_{0.99} \approx [1.4; 2.3].$$

Teisi võimalikke vahemikhinnangu leidmise viise käsitleme järgmises osas.

1.3. Statistiliste hüpoteeside kontroll

Asume õpiku ühe olulisema teema juurde, mis käsitleb statistilisi hüpoteese ning nende kontrolle. Hüpoteeside kontroll on matemaatilise statistika põhieesmärk. Just selle põhjal interpreteeritakse enamik teaduslike

eksperimentide tulemusi. Sissejuhatuseks defineerime mõiste statistiline hüpotees.

Definitsioon 1.17. Statistiliseks hüpoteesiks nimetatakse oletust juhusliku suuruse jaotuse või selle jaotuse parameetri kohta.

1.3.1. Hüpoteeside kontrolli ideoloogiline olemus

Anname esmalt ülevaate statistiliste hüpoteeside põhimõtetest. Statistilise hüpoteesi eesmärk on kummutada varem teada olevat tõde. Seda tõde nimetatakse nullhüpoteesiks.

Definitsioon 1.18. Nullhüpoteesiks nimetakse oletust, mida arvatakse olevat tõene enne statistilist testi.

Nullhüpoteesi tähistatakse kui H_0. Nullhüpoteesile alternatiivset tõde nimetatakse sisukaks hüpoteesiks.

Definitsioon 1.19. Sisukaks hüpoteesiks nimetatakse oletust, mida me tahame tõestada statistilise testiga.

Sisukat hüpoteesi tähistakse kui H_1. Hüpoteesid H_0 ja H_1 moodustavad hüpoteeside paari, mis formuleeritakse kui

$$\begin{cases} H_0: \text{ väide}, \\ H_1: \text{ vastuväide}. \end{cases}$$

Nullhüpotees ja sisukas hüpotees tuleb sõnastada selliselt, et kehtiksid järgmised seosed:

$$P(H_0 \cup H_1) = 1 \text{ ja } P(H_0 \cap H_1) = 0.$$

Järelikult kehtib parajasti üks hüpoteesidest H_0 või H_1 ehk väide H_1 on väite H_0 eitus.

Nullhüpoteesid erinevates valdkondades võiksid olla järgmised:

õigusteaduses H_0 : süütuse presumptsioon;

meditsiinis H_0 : sünteesitud keemiline ühend ei ole ravim;

keskkonnateadustes H_0 : jõgede reostuskoormused ei ole muutunud.

Üldiselt tähendab nullhüpotees H_0 endist seisukohta ehk *status quo* olekut. Sisuka hüpoteesi H_1 näol on tegemist muutusega ehk uue avastusega.

Kuna hüpoteeside kontroll põhineb juhuslikul suurusel, siis teeme selle juures paratamatult vigu. Neid on kahte liiki: I liiki viga ja II liiki viga.

Definitsioon 1.20. Riski lugeda hüpotees H_1 tõestatuks tingimusel, et on õige nullhüpotees H_0 nimetatakse I liiki veaks.

Definitsioon 1.21. Riski jääda hüpoteesi H_0 juurde tingimusel, et on õige sisukas hüpotees H_1 nimetatakse II liiki veaks.

Tähistagem I ja II liiki viga vastavalt kui γ_1 ja γ_2. Seega

$$\gamma_1 = P(\text{tõestada } \ H_1 \mid \text{õige on } \ H_0)$$

ja

$$\gamma_2 = P(\text{jääda } \ H_0 \ \text{ juurde} \mid \text{õige on } \ H_1).$$

Hüpoteeside kontrollil tehtavad vead võtab kokku alljärgnev tabel:

On tegelikult / Arvame olevat	Nullhüpotees H_0	Sisukas hüpotees H_1
Nullhüpotees H_0	Õige	II liiki viga γ_2
Sisukas hüpotees H_1	I liiki viga γ_1	Õige

Tekib kohe küsimus, kumb vigadest on halvem, kas I või II liiki. Vastus sellele küsimusele on, et I liiki viga on palju halvem kui II liiki viga. Seega tuleb statistiliste hüpoteeside kontrollimisel lähtuda kohtumõistmise põhimõttest, *et ennem jäägu süüdlane karistamata, kui süütu süüdi mõista.* Seoses riskiga teha I liiki viga, tuleb sisse mõiste olulisustõenäosus.

Definitsioon 1.22. Olulisustõenäosuseks nimetatakse valimi realisatsiooni põhjal leitud riski teha I liiki viga.

Definitsioonis 1.22 on oluline rõhutada fraasi *valimi realisatsiooni põhjal*, sest olulisustõenäosusest ei saa teha järeldusi tegeliku I liiki vea tegemise

riskist. Selle järelduse usaldusväärsus sõltub sellest, kuivõrd hästi kirjeldab valim üldkogumit. Üldjuhul on I liiki vea tegemise risk suurem kui olulisustõenäosus.

Olulisustõenäosust tähistatakse enamasti ingliskeelse terminiga *p-value*. Leidmaks suuruse *p-value* väärtus tuleb koostada valimi põhjal teststatistik $T(\mathbf{X})$. See koostatakse eeldusel, et nullhüpotees H_0 on õige. Enamasti on teststatistik pidev juhuslik suurus. Olgu mõõtmistulemuste põhjal saadud teststatistiku väärtus $T(\mathbf{x}) = t$. Püstitame küsimuse: kui suur on tõenäosus saada teststatistiku väärtuseks t, kui H_0 on õige. Küsitav tõenäosus ongi olulisustõenäosus (*p-value*). Selle leidmiseks määratakse väärtusele t vastav piirkond $\mathcal{H}_t$. See piirkond sõltub testi olemusest ja teststatistikule vastavast tihedusfunktsioonist. Üldiselt leitakse piirkond $\mathcal{H}_t$ järgmiselt.

1) Kui see tihedusfunktsioon on paarisfunktsioon (ehk tema graafik on sümmeetriline), siis $\mathcal{H}_t = (-\infty; -t)$, $\mathcal{H}_t = (t; \infty)$ või $\mathcal{H}_t = (-\infty; -t) \cup (t; \infty)$.

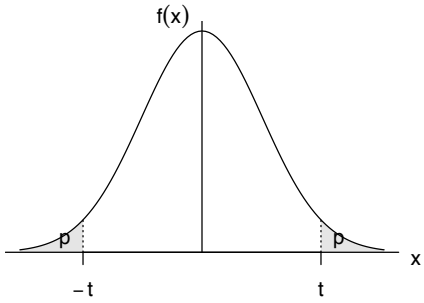
2) Kui teststatistiku väärtus on positiivne ja tema tihedusfunktsiooni graafik ei ole sümmeetriline, siis $\mathcal{H}_t = (0; t)$ või $\mathcal{H}_t = (t; \infty)$.

Esitame olulisustõenäosuse piirkonna $\mathcal{H}_t$ kaudu. Selle piirkonna abil saame tingliku tõenäosuse

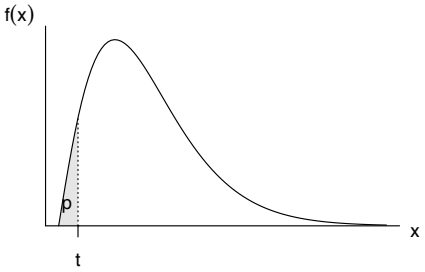
$$p\text{-}value = P(T(\mathbf{X}) \in \mathcal{H}_t \mid \text{õige on } \ H_0). \tag{1.8}$$

Seoses (1.8) peitub *p-value* tõeline olemus. Sellest tulenevalt on olulisustõenäosus tinglik tõenäosus tingimusel, et nullhüpotees H_0 on õige. Teisisõnu, kui suur on tõenäosus saada valimi antud realisatsiooni, kui kehtib nullhüpotees? Näiteks: kui suure tõenäosusega saadi 44 korral 100 mündiviskest vapp, kui eeldati mündi sümmeetrilisust?

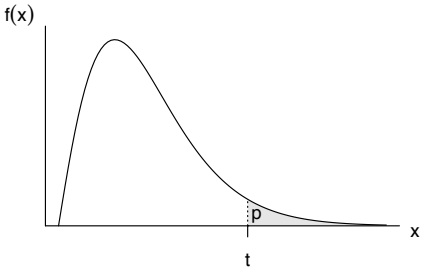
Alljärgnevad graafikud illustreerivad olulisustõenäosusi sümmeetrilisel (joonis 1.4) ja mittesümmeetrilisel (joonised 1.5–1.6) juhul.



Joonis 1.4. Piirkonnale $\mathcal{H}_t = (-\infty; -t) \cup (t; \infty)$ vastav olulisustõenäosus p



Joonis 1.5. Piirkonnale $\mathcal{H}_t = (0; t)$ vastav olulisustõenäosus p



Joonis 1.6. Piirkonnale $\mathcal{H}_t = (t; \infty)$ vastav olulisustõenäosus p

Paratamatult kerkib üles küsimus, et maksimaalselt kui suurt olulisustõenäosust võime endale lubada. Selle tõenäosuse määrab meile olulisuse nivoo (ingl *significance level*).

Definitsioon 1.23. Olulisuse nivooks nimetatakse maksimaalset lubatud *p-value* väärtust.

Tähistagem olulisuse nivood kui β. Olulisuse nivoo β ja usaldusnivoo α avalduvad üksteise kaudu kui

$$\beta = 1 - \alpha.$$

Otsustuskriteerium valimaks kas nullhüpoteesi või sisukat hüpoteesi on järgmine:

1) kui $p\text{-}value < \beta$, siis loeme tõestatuks sisuka hüpoteesi H_1;

2) kui $p\text{-}value \geq \beta$, siis oleme sunnitud jääma nullhüpoteesi H_0 juurde.

Milline aga on parim olulisuse nivoo β? See sõltub uurimise valdkonnast, mille kohta tehakse statistilist hüpoteesi. Need valdkonnad võib jagada tinglikult 3 gruppi.

1) Suhteliselt vigadealdistes valdkondades, nagu sotsioloogilised küsitlused, võetakse olulisuse nivooks $\beta = 0.1$.

2) Loodus- ja insneriteadustes sobib, kui võtta $\beta = 0.05$.

3) Testides, mille otsustustest sõltub inimsaatus või inimelu, tuleb olulisuse nivooks β võtta 0.01. Säärased valdkonnad on näiteks õigusteadus ja meditsiin.

Enamasti kuulub olulisuse nivoo väärtus lõiku $[0.01; 0.1]$.

Vaatame järgnevalt statistilise testimise n-ö kaudset meetodit, mis põhineb teststatistikule kriitilise väärtuse (ingl *critical value*) leidmisel. Tähistagem seda väärtust kui t_{crit}. Kui teststatistiku tihedusfunktsiooni graafik ei ole sümmeetriline, siis võib olla kriitilise väärtusi kaks: t_{crit_1} ja t_{crit_2}. Olgu $t_{crit_1} < t_{crit_2}$. Teststatistiku kriitiline väärtus või kriitilised väärtused valitakse selliselt, et talle vastaks olulisustõenäosus, mis on võrdne olulisuse nivooga β. See tähendab

$$P(T(\mathbf{X}) \in \mathcal{H}_{crit} \mid \text{õige on } \ H_0) = \beta,$$

kus kriitiline piirkond $\mathcal{H}_{crit}$ sõltub väärtusest t_{crit} või väärtustest t_{crit_1} ja (või) t_{crit_2}.

Kriitilise väärtuse alusel jagatakse teststatistiku väärtused 2 piirkonda:

1) nullhüpoteesi piirkonda $\mathcal{H}_0$,

2) sisuka hüpoteesi piirkonda $\mathcal{H}_1$.

Piirkondadeks $\mathcal{H}_0$ ja $\mathcal{H}_1$ jagamisel kehtib põhimõte, kus

$$P(T(\mathbf{X}) \in \mathcal{H}_1) = \beta \text{ ja } P(T(\mathbf{X}) \in \mathcal{H}_0) = 1 - \beta$$

ehk

$$\mathcal{H}_1 = \mathcal{H}_{crit} \text{ ja } \mathcal{H}_0 = \mathcal{H}_{crit}^c.$$

Piirkondade $\mathcal{H}_0$ ja $\mathcal{H}_1$ määramisel kasutatakse peamiselt kahte varianti:

1) kui teststatistiku tihedusfunktsioon on paarisfunktsioon, siis

$$\mathcal{H}_0 = [-t_{crit}; 0], \quad \mathcal{H}_0 = [-t_{crit}; t_{crit}] \text{ või } \mathcal{H}_0 = [0; t_{crit}]$$

ning

$$\mathcal{H}_1 = (-\infty; -t_{crit}), \quad \mathcal{H}_1 = (t_{crit}; \infty) \text{ või } \mathcal{H}_1 = (-\infty; -t_{crit}) \cup (t_{crit}; \infty);$$

2) teststatistiku tihedusfunktsiooni mittesümmeetrilise graafiku korral

$$\mathcal{H}_0 = [t_{crit_1}; t_{crit_2}], \quad \mathcal{H}_0 = [t_{crit_1}; \infty] \text{ või } \mathcal{H}_0 = [0; t_{crit_2}]$$

ning

$$\mathcal{H}_1 = [0; t_{crit_1}), \quad \mathcal{H}_1 = (t_{crit_2}; \infty) \text{ või } \mathcal{H}_1 = [0; t_{crit_1}) \cup (t_{crit_2}; \infty).$$

Kui valimi põhjal leitud teststatistiku väärtus t kuulub piirkonda $\mathcal{H}_1$, siis loeme tõestatuks sisuka hüpoteesi H_1. Kui aga piirkonda $\mathcal{H}_0$, siis oleme kohustatud jääma nullhüpoteesi H_0 juurde.

1.3.2. Statistilise testi võimsus

Statistiliste hüpoteeside valik on sageli väga lai. Paratamatult tekib küsimus, milline test on mingile kindlale teadusprobleemile parim. Üheks testi headuse kriteeriumiks on võimsusfunktsioon. Me tahame teststatistikuga $T(\mathbf{X})$ testida jaotuse parameetri θ_j väärtust. Sel juhul saame defineerida selle testi võimsusfunktsiooni alljärgnevalt.

Definitsioon 1.24. Statistilise testi võimsusfunktsiooniks $h(\theta_j)$ nimetatakse sisuka hüpoteesi tõestamise tõenäosust tingimusel, et väärtus θ_j on üldkogumi õige väärtus:

$$h(\theta_j) = P(\text{tõestada } \ H_1 \mid \theta_j).$$

Seega eeldatakse, et antud valimi põhjal leitud jaotuse parameetri väärtus on õige. Võimsusfunktsiooni abil saame leida tõenäosused sisuka hüpoteesi tõestamiseks parameetri θ_j erinevate väärtuste korral. Seda funktsiooni saab avaldada II liiki vea tegemise tõenäosuse kaudu järgmiselt:

$$h(\theta_j) = 1 - P(\text{jääda } \ H_0 \ \text{ juurde} \mid \text{õige on } \ H_1) =$$

$$= 1 - P(\text{me teeme II liiki vea}) = 1 - \gamma_2.$$

Tõenäosust $1-\gamma_2$ nimetatakse statistilise testi võimsuseks. Uurime järgnevalt paari näite abil võimsusfunktsiooni konstrueerimist.

Näide 1.15. Olgu meil valim, mille element $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Esitame järgmise hüpoteeside paari normaaljaotuse keskväärtusele μ :

$$\begin{cases} H_0 : \mu = 10, \\ H_1 : \mu \neq 10. \end{cases}$$

Olgu standardhälve $\sigma = 5$. Siis sobib teststatistikuks

$$Z = \frac{\overline{x} - 10}{5}\sqrt{n} \sim \mathcal{N}(0, 1),$$

kus $\overline{x}$ tähistab valimi põhjal leitud aritmeetilist keskmist. Võtame endale järgmise otsustuskriteeriumi:

1) kui $\mu_1 \leq \overline{x} \leq \mu_2$, siis jääme nullhüpoteesi H_0 juurde;

2) kui aga $\overline{x} < \mu_1$ või $\overline{x} > \mu_2$, siis loeme tõestatuks sisuka hüpoteesi H_1.

Selliste otsustuskriteerimi piiride korral saame testi võimsusfunktsiooniks

$$h(\mu) = P(\text{tõestada } \ H_1 \mid \mu) = 1 - \gamma_2 = 1 - P(\mu_1 \leq \overline{x} \leq \mu_2 \mid \mu) =$$

$$= 1 - \left\{ \Phi\left(\frac{\mu_2 - \mu}{5}\sqrt{n}\right) - \Phi\left(\frac{\mu_1 - \mu}{5}\sqrt{n}\right) \right\} =$$

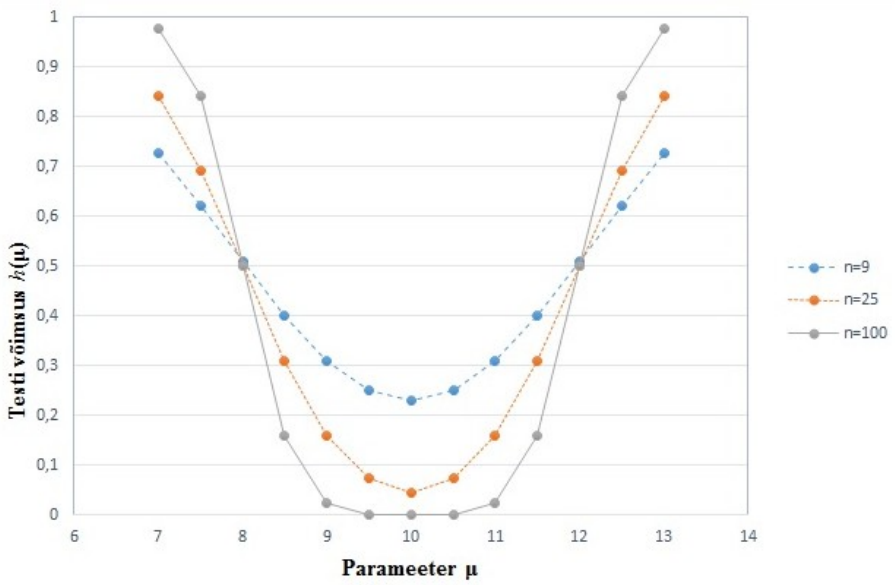
$$= 1 + \Phi\left(\frac{\mu_1 - \mu}{5}\sqrt{n}\right) - \Phi\left(\frac{\mu_2 - \mu}{5}\sqrt{n}\right).$$

Olgu $\mu = 11$, $\mu_1 = 8$, $\mu_2 = 12$ ning $n = 9$. Siis

$$h(11) = 1 + \Phi\left(\frac{3(8-11)}{5}\right) - \Phi\left(\frac{3(12-11)}{5}\right) =$$

$$= 1 + \Phi(-1.8) - \Phi(0.6) = 1 - 0.464 - 0.226 = 0.31.$$

Kui aga $\mu = 10$, siis $h(\mu) \approx 0.23$. Ideaaljuhul aga peab $h(10) = 0$. Allpool on toodud võimsusfunktsiooni graafikud valimi mahtude $n = 9$, $n = 25$ ja $n = 100$ ning otsustuskriteeriumi piiride $\mu_1 = 8$ ja $\mu_2 = 12$ korral.



Joonis 1.7. Statistilise testi võimsus erinevate valimi mahtude n korral

Võimsaima testi korral ehk $n \to \infty$ avaldub antud otsustuskriteeriumi puhul võimsusfunktsioon järgmiselt:

$$h(\mu) = \begin{cases} 0, \mu \in [8;12], \\ 1, \mu \notin [8;12]. \end{cases}$$

Ideaaljuhul ehk juhul, kui $n \to \infty$ ja $\mu_2 - \mu_1 \to 0$ saame võimsusfunktsiooniks

$$h(\mu) = \begin{cases} 1, \text{ kui } \mu \neq 10, \\ 0, \text{ kui } \mu = 10. \end{cases}$$

Näide 1.16. Olgu meil nüüd normaaljaotusega juhusliku suuruse keskväärtusele μ hüpoteesid

$$\begin{cases} H_0 : \mu \geq 10, \\ H_1 : \mu < 10. \end{cases}$$

Olgu olulisuse nivoo $\beta = 0.05$. Siis on otsustuskriteerium järgmine:

1) kui $\dfrac{\mu-10}{\sigma}\sqrt{n} < -1.64$, siis loeme tõestatuks sisuka hüpoteesi H_1;

2) kui $\dfrac{\mu-10}{\sigma}\sqrt{n} \geq -1.64$, siis jääme nullhüpoteesi H_0 juurde.

Sellele kriteeriumile vastav võimsusfunktsioon

$$h(\mu) = 1-\gamma_2 = 1 - P\left(\frac{\mu-10}{\sigma}\sqrt{n} \geq -1.64\right) =$$

$$= 1 - P\left(\frac{10-\mu}{\sigma}\sqrt{n} \leq 1.64\right) = 0.5 - \Phi\left(1.64 - \frac{10-\mu}{\sigma}\sqrt{n}\right).$$

Kui $\sigma = 5$ ning $n = 9$, siis

$$h(\mu) = 0.5 - \Phi(0.6\mu - 4.36).$$

Olulisuse nivoost β sõltub statistilise testi võimsus sellisel juhul järgmiselt:

$$h(\mu) = 0.5 - \Phi(z_{1-\beta} - 6 + 0.6\mu).$$

Võimsaima testi korral

$$h(\mu) = \begin{cases} 1, & \text{kui } \mu < 10, \\ 0, & \text{kui } \mu \geq 10. \end{cases}$$

Kokkuvõttes saab öelda, et statistilise testi võimsus sõltub peamiselt:

1) valimi mahust n,

2) mõõtmise täpsusest ehk näites 1.15 kirjeldatud lõigu $[\mu_1; \mu_2]$ pikkusest,

3) olulisuse nivoost β.

Mida suurem on valimi maht ning mida täpsem on mõõtmine, seda võimsama testi saame.

1.3.3. Näiteid erinevatest statistilistest hüpoteesidest

Statistiliste hüpoteeside kontrollimisel tuleb teada teststatistiku $T(\mathbf{X})$ jaotust. Selle jaotuse põhjal võib statistilised hüpoteesid jagada 4 gruppi. Uurime neid gruppe lähemalt, tehes sellega ühtlasi tutvust teststatistikute jaotustega.

Normaalne aproksimatsioon

Normaalne aproksimatsioon ehk normaaljaotusega lähendamine töötab juhul, kui valimi maht $n \geq 30$. See meetod põhineb tsentraalsel piirteoreemil, millest tuleneb teststatistik

$$Z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1).$$

Edaspidi nimetame normaalsele aproksimatsioonile vastavat statistikut Z-statistikuks. Vaatleme lähemalt 3 hüpoteeside paari.

Kahepoolne hüpotees

Kui tegemist on kahepoolse hüpoteesiga, siis on hüpoteeside paar järgmine:

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{cases}$$

Leiame antud hüpoteeside paarile vastava statistiku kriitilise väärtuse z_{crit}. Saame, et olulisuse nivoo

$$\beta = P(\mid Z \mid > z_{crit}) = 1 - P(\mid Z \mid \leq z_{crit}) =$$

$$= 1 - P(-z_{crit} \leq Z \leq z_{crit}) = 1 - 2\Phi(z_{crit}),$$

millest järeldub, et

$$z_{crit} = \Phi^{-1}\left(\frac{1-\beta}{2}\right).$$

Otsustuseeskiri on järgmine: kui valimi põhjal saadud statistiku Z väärtus $z \leq z_{crit}$, peame jääma nullhüpoteesi H_0 juurde, kui aga $z > z_{crit}$, siis loeme tõestatuks sisuka hüpoteesi H_1. Esitame alljärgnevas tabelis olulisuse nivoole β vastavad kriitilised väärtused:

β	z_{crit}
0.1	1.64
0.05	1.96
0.01	2.58

Teststatistiku väärtusele z vastav olulisustõenäosus

$$p\text{-}value = 2(0.5 - \Phi(\mid z \mid)).$$

Vasakpoolne hüpotees

Hüpoteeside paar on vasakpoolse hüpoteesi korral järgmine:

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu < \mu_0. \end{cases}$$

Antud juhul olulisuse nivoo

$$\beta = P(Z < -z_{crit}) = 0.5 + \Phi(-z_{crit}),$$

millest järeldub, et

$$-z_{crit} = \Phi^{-1}(\beta - 0.5).$$

Sisukas hüpotees H_1 loetakse tõestatuks, kui valimi põhjal saadud $z < -z_{crit}$. Vastasel korral peame jääma nullhüpoteesi H_0 juurde. Antud juhul saame teststatistiku väärtusele z vastava olulisustõenäosuse

$$p\text{-}value = \Phi(z) + 0.5.$$

Parempoolne hüpotees

Parempoolsele hüpoteesile vastab hüpoteeside paar

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu > \mu_0. \end{cases}$$

Antud juhul olulisuse nivoo

$$\beta = P(Z > z_{crit}) = 1 - P(Z \leq z_{crit}) = 1 - (0.5 + \Phi(z_{crit})),$$

millest järeldub, et

$$z_{crit} = \Phi^{-1}(0.5 - \beta).$$

Sisukas hüpotees H_1 loetakse tõestatuks, kui valimi põhjal saadud $z > z_{crit}$. Vastasel korral peame jääma nullhüpoteesi H_0 juurde. Antud juhul saame teststatistiku väärtusele z vastava olulisustõenäosuse

$$p\text{-}value = 0.5 - \Phi(z).$$

Rakendame erinevatele jaotustele normaalset aproksimatsiooni. Üldisust kitsendamata teeme seda kahepoolsetele hüpoteesidele.

Rakendused binoomjaotusele

Antud juhul me katsetame n korda sündmuse A toimumist. Eesmärk on uurida selle sündmuse tõenäosust p. Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)$, kus

$$X_i = \begin{cases} 1, \text{ kui toimub } A, \\ 0, \text{ kui toimub } \overline{A} \end{cases}$$

ning hüpoteeside paar

$$\begin{cases} H_0 : p = p_0, \\ H_1 : p \neq p_0. \end{cases}$$

Antud juhul on nullhüpoteesi korral $E(X_i) = p_0$ ja $D(X_i) = p_0(1 - p_0)$. Seega vastav teststatistik

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n},$$

kus $\widehat{p} = \dfrac{k}{n}$, milles k tähistab jaatajate hulka.

Näide 1.17. Üks erakond sai valimistel 22.5% häältest. Vahetult enne valimisi küsitleti 100 inimest, kellest 28 pooldas seda erakonda. Kas võib olulisuse nivool 0.1 väita, et küsitlus hindas erakonna populaarsust üle?

Olgu p erakonna toetajate osakaal. Koostati järgmine hüpoteeside paar:

$$\begin{cases} H_0 : p = 0.225, \\ H_1 : p > 0.225. \end{cases}$$

Selle paari kontrollimiseks koostati teststatistik

$$Z = \frac{\widehat{p} - 0.225}{\sqrt{0.225(1-0.225)}}\sqrt{100} \sim \mathcal{N}(0,1).$$

Küsitletute hulk 100 on piisavalt suur eeldamaks tsentraalse piirteoreemi toimimist. Küsitluste põhjal saadi $\widehat{p} = \dfrac{28}{100}$. Teststatistiku väärtuseks saadi, et

$$z = \frac{0.28 - 0.225}{\sqrt{0.225 \cdot 0.775}}10 \approx 1.32.$$

Leiame saadud väärtusele vastva olulisustõenäosuse. Kuna tegemist on parempoolse hüpoteesiga, siis

$$p\text{-}value = 0,5 - \Phi(1.32) = 0.5 - 0.407 = 0.093 < 0.1.$$

Kuna olulisustõenäosus osutus väiksemaks kui olulisuse nivoo 0.1, siis loeme tõestatuks sisuka hüpoteesi H_1. Seega võime väita, et küsitlus tõepoolest hindas erakonna populaarsust üle. Kuid väga napilt, sest kriitiline väärtus $z_{crit} = \Phi^{-1}(0.5 - 0.1) \approx 1.28$. Antud juhul saame järgmised piirkonnad:

$$\mathcal{H}_t = (1.32; \infty), \quad \mathcal{H}_1 = (1.28; \infty) \text{ ja } \mathcal{H}_0 = [0; 1.28].$$

Olgu meil kaks valimit mahtudega vastavalt n_1 ja n_2. Mõlemad mahud tähendavad sõltumatute katsete hulka katsetamaks sündmuse A toimumist. Toimugu see sündmus esimese valimi juhul x_1 ja teise valimi juhul x_2 korda. Seega vastavad juhuslikud $X_1 \sim B(n_1, p_1)$ ja $X_2 \sim B(n_2, p_2)$. Me tahame testida, kas sündmuse A toimumise tõenäosused p_1 ja p_2 on võrdsed. Selleks koostame hüpoteeside paari

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2. \end{cases}$$

Uurime statistikut $\widehat{p}_1 - \widehat{p}_2 = \dfrac{x_1}{n_1} - \dfrac{x_2}{n_2}$. Saame, et keskväärtus

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2.$$

Arvestades klassikalise statistika sõltumatuse eeldust, saame dispersiooniks

$$D(\widehat{p}_1 - \widehat{p}_2) = D(\widehat{p}_1) + D(\widehat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Nullhüpoteesi H_0 kehtimise korral saame statistikuks

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \frac{\widehat{p}_1 - \widehat{p}_2}{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

kus p tähistab ühist osakaalu (nullhüpoteesi korral $p = p_1 = p_2$). Selle osakaalu STP hinnang $\widehat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$. Suure valimi mahu $n_1 + n_2$ korral allub statistik

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

ligikaudu standardsele normaaljaotusele. Selle statistiku põhjal kontrollitakse hüpoteesi tõenäosuste p_1 ja p_2 võrdsusest.

Rakendused Poissoni jaotusele

Antud juhul on meil valim, mille element $X_i =\sim Po(\lambda)$ $i = 1, 2, ..., n$. Uurime mingi (haruldase) sündmuse esinemise sagedust ning testime Poissoni jaotuse parameetrit λ. Hüpoteeside paar on nüüd järgmine:

$$\begin{cases} H_0 : \lambda = \lambda_0, \\ H_1 : \lambda \neq \lambda_0. \end{cases}$$

Nullhüpoteesi H_0 korral $E(X_i) = D(X_i) = \lambda_0$. Vastavalt parameetri λ STP hinnangule saame teststatistikuks

$$Z = \frac{\overline{x} - \lambda_0}{\sqrt{\lambda_0}}\sqrt{n}.$$

Rakendused eksponentjaotusele

Antud juhul on meil valim, mille element $X_i \sim \mathcal{E}(\nu)$, $i = 1, 2, ..., n$, $\nu > 0$. Olgu nullhüpoteesi korral $\nu = \nu_0$. Seega saame järgmise hüpoteeside paari keskväärtusele μ :

$$\begin{cases} H_0 : \mu = \dfrac{1}{\nu_0}, \\ H_1 : \mu \neq \dfrac{1}{\nu_0}. \end{cases}$$

Kuna nullhüpoteesi H_0 korral $E(X_i) = \sqrt{D(X_i)} = \dfrac{1}{\nu_0}$, siis arvestades STP hinnangut parameetrile ν, saame teststatistikuks

$$Z = \frac{\overline{x} - \dfrac{1}{\nu_0}}{\dfrac{1}{\nu_0}} \sqrt{n}.$$

Hii-ruut testid

Defineerime jaotuse, mida nimetatakse χ^2-jaotuseks (loe hii-ruut). Kuigi alljärgnevat on matemaatilises mõttes korrektsem nimetada kas lauseks või teoreemiks, sõnastame selle kui definitsiooni.

Definitsioon 1.25. Olgu meil sõltumatud juhuslikud suurused

$$X_1, X_2, ..., X_n.$$

Olgu $X_i \sim \mathcal{N}(0,1)$, $i = 1, 2, ..., n$. Siis juhuslik suurus

$$Y_n = \sum_{i=1}^{n} X_i^2$$

on χ^2-jaotusega vabadusastmete arvuga n.

Juhusliku suuruse Y_n allumist χ^2-jaotusele vabadusastmete arvuga n tähistagem kui $Y_n \sim \chi^2(n)$. Huviline võib χ^2-jaotusest ja tema tihedusfunktsioonist lugeda õpikust [41] (lk 137–139). Selgitus – χ^2-jaotusele vastava tihedusfunktsiooni graafik ei ole sümmeetriline. Sellele jaotusele vastav keskväärtus ja dispersioon avalduvad järgmiselt:

$$E(Y_n) = n \text{ ning } D(Y_n) = 2n.$$

Kui vabadusastme arv $n = 2$, siis on χ^2-jaotus identne eksponentjaotusega, mille parameeter on $\frac{1}{2}$. Seega on χ^2-jaotuse näol tegemist eksponentjaotuse üldistamisega.

Sõnastame järgnevalt kaks χ^2-jaotust iseloomustavat tulemust. Need tulemused leiavad hiljem rakendust χ^2-jaotusele alluvate teststatistikute moodustamisel.

Lause 1.2. Olgu meil sõltumatud juhuslikud suurused $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Olgu

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Siis juhuslik suurus

$$H = \frac{\sum_{i=1}^{n} (X_i - \overline{x})^2}{\sigma^2}$$

on χ^2-jaotusega vabadusastmete arvuga $n - 1$.

Lause 1.3. Olgu juhuslikud suurused $X_1 \sim \chi^2(k_1)$, $X_2 \sim \chi^2(k_2)$, ..., $X_n \sim \chi^2(k_n)$ sõltumatud. Siis

$$H = \sum_{i=1}^{n} X_i^2 \sim \chi^2\left(\sum_{i=1}^{n} k_i \right).$$

Lause 1.3 väidab, et χ^2-jaotus on aditiivne.

Uurime järgnevalt χ^2-teste erinevate statistiliste hüpoteeside kontrollimisel.

Normaaljaotusega juhusliku suuruse disperiooni test

Definitsioonist 1.25 ja lausest 1.2 järeldub, et

$$H = \frac{s^2}{\sigma^2}(n-1) \sim \chi^2(n-1),$$

kui $X_i \sim \mathcal{N}(\mu, \sigma)$.

Püstitame normaaljaotusega juhusliku suuruse dispersiooni kohta hüpoteeside paari

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2, \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{cases}$$

Seda saab kontrollida teststatistikuga H, mis nullhüpoteesi H_0 kehtivuse korral on järgmine:

$$H = \frac{s^2(n-1)}{\sigma_0^2}.$$

Statistiku H põhjal saame konstrueerida normaaljaotusega juhusliku suuruse dispersioonile α-usaldusintervalli. Tähistades selle statistiku α-kvantiili vabadusastmete arvu n korral kui $h_{\alpha;n}$, saame, et

$$P\left(h_{\frac{1-\alpha}{2};n-1} \leq \frac{s^2}{\sigma^2}(n-1) \leq h_{\frac{1+\alpha}{2};n-1}\right) = \alpha \iff$$

$$\iff P\left(\frac{h_{\frac{1-\alpha}{2};n-1}}{s^2(n-1)} \leq \frac{1}{\sigma^2} \leq \frac{h_{\frac{1+\alpha}{2};n-1}}{s^2(n-1)}\right) = \alpha \iff$$

$$\iff P\left(\frac{s^2(n-1)}{h_{\frac{1+\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{s^2(n-1)}{h_{\frac{1-\alpha}{2};n-1}}\right) = \alpha.$$

Saime α-usaldusintervalli

$$I_\alpha = \left[\frac{s^2(n-1)}{h_{\frac{1+\alpha}{2};n-1}}; \frac{s^2(n-1)}{h_{\frac{1-\alpha}{2};n-1}}\right].$$

Näide 1.18. Olgu meil juhuslik suurus $X \sim \mathcal{N}(\mu, \sigma)$. Dispersiooni σ^2 nihketa hinnanguks saadi 10-elemendilise valimi põhjal s^2. Siis dispersiooni 0.95-usaldusintervall

$$I_{0.95} = \left[\frac{9s^2}{19.02}; \frac{9s^2}{2.7}\right].$$

Jaotuse sobimise test

Inglise keeles on selle testi nimi *Goodness of Fit Test*. Selle testiga kontrollitakse, kas uuritav suurus allub eeldatavale jaotusele. Me tahame ümber lükata varasemast tuntud tõde, et uuritav suurus allub teada olevale jaotusele, nagu on näiteks järgmistes olukordades:

1) täringuviskel on kõikide silmade saamise tõenäosused $\frac{1}{6}$;

2) lambipirni või patarei eluiga allub eksponentjaotusele;

3) kehakaal ja kehapikkus alluvad normaaljaotusele.

Hindamaks tegeliku jaotuse sarnasust teoreetilisega koostatakse mõõtmistulemuste põhjal sagedustabel

Klass	Sagedus
$[x_1; x_2)$	n_1
$[x_2; x_3)$	n_2
$\vdots$	$\vdots$
$[x_i; x_{i+1})$	n_i
$\vdots$	$\vdots$
$[x_m; x_{m+1})$	n_m

Tabelis on uuritav suurus jagatud suurusklassidesse, kus n_i tähistab vastavasse klassi kuuluvate vaatluste hulka. Iga vaatlus kuulub parajasti ühte klassi ning kehtib seos

$$\sum_{i=1}^{m} n_i = n.$$

Püstitatakse järgmine hüpoteeside paar:

$$\begin{cases} H_0 : X \sim F, \\ H_1 : X \text{ on muu jaotusega .} \end{cases}$$

Kontrollimaks seda hüpoteeside paari koostatakse teststatistik

$$H = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{1.9}$$

kus

$$p_i = P(x_i \le X \le x_{i+1}) = F(x_{i+1}) - F(x_i).$$

Statistikut H nimetatakse Pearsoni χ^2-ruut kriteeriumiks. Millisele jaotusele allub see statistik?

Lause 1.4. Seosega (1.9) defineeritud teststatistik allub ligikaudu χ^2-jaotusele vabadusastmete arvuga $m - k - 1$, kus k tähistab jaotuse F parameetrite hulka.

Tõestus. Esitame antud väite tõestuse skeemi. See tõestus põhineb kolmel tulemusel:

1) Poissoni piirteoreemil,

2) tsentraalsel piirteoreemil,

3) lausel 1.3 ning lause 1.2 üldistusel.

Suurus p_i on i-ndasse klassi kuulumise tõenäosus, $i = 1, 2, ..., m$. Seda kuulumist katsetatakse sõltumatult n korda. Seega on n_i juhuslik suurus, mis allub binoomjaotusele keskväärtusega np_i. Vastavalt tsentraalsele piirteoreemile on juhuslik suurus

$$Z_i = \frac{n_i - np_i}{\sqrt{np_i(1-p_i)}},$$

$i = 1, 2, ..., m$, ligikaudu standardse normaaljaotusega. Poissoni piirteoreemi (mille tõestuse võib leida õpikust [34], lk 75–76) kohaselt läheneb Z_i jaotus Poissoni jaotusele, kui $n \to \infty$. Poissoni jaotuse puhul aga

$$E(n_i) = D(n_i) = np_i.$$

Tsentraalse piirteoreemi ning Poissoni piirteoreemi toimimisel järeldub definitsioonist 1.25 ning lausetest 1.2 ja 1.3, et juhusliku suuruse

$$H = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

jaotus läheneb χ^2-jaotusele vabadusastmete arvuga $m - k - 1$.

□

Kui hüpoteesi H_0 väites olev jaotus F on normaaljaotus, siis $k = 2$, eksponentjaotuse ja Poissoni jaotuse puhul $k = 1$.

Tähistame seoses (1.9) suuruse n_i kui emp_i ning suuruse np_i kui teor_i. Siis statistik

$$H = \sum_{i=1}^{m} \frac{(\text{emp}_i - \text{teor}_i)^2}{\text{teor}_i}.$$

Seega kujutab Pearsoni χ^2-ruut kriteeriumi iga liidetav endast empiirilise sageduse (emp_i) ning teoreetilise sageduse (teor_i) vahe ruudu ja teoreetilise sageduse suhet. Teoreetilise sageduse all on mõeldud sagedust nullhüpoteesi kehtimise korral.

Lause 1.4 väide hakkab kehtima, kui empiirilised ning teoreetilised sagedused on piisavalt suured. Rusikareegliks on nõue, et $n_i \geq 5$ ning $np_i \geq 5$. Kui mõningasse klassi langemise sagedus on liiga väike, siis liidetakse ta naaberintervalli(de)ga kuni tingimuse $n_i \geq 5$ täitumiseni.

Sõltumatuse test

Kontrollime, kas kaks juhuslikku suurust on sõltumatud. Seda tehakse sagedustabelite abil. Olgu meil juhuslikud suurused $X = X_1, X_2, ..., X_l$ ja $Y = Y_1, Y_2, ..., Y_r$. Tähistagu sümbol $\perp$ sõltumatust. Kontrollime hüpoteeside paari

$$\begin{cases} H_0 : X \perp Y, \\ H_1 : X, Y \text{ ei ole sõltumatud.} \end{cases}$$

Kanname mõõtmistulemused sagedustabelisse:

X \ Y	Y_1	$\cdots$	Y_j	$\cdots$	Y_r	X sagedus
X_1	k_{11}	$\cdots$	k_{1j}	$\cdots$	k_{1r}	$k_{1.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
X_i	k_{i1}	$\cdots$	k_{ij}	$\cdots$	k_{ir}	$k_{i.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
X_l	k_{l1}	$\cdots$	k_{lj}	$\cdots$	k_{lr}	$k_{l.}$
Y sagedus	$k_{.1}$	$\cdots$	$k_{.j}$	$\cdots$	$k_{.r}$	$\Sigma = n$

Suurus k_{ij} tähistab sagedustabelis selliste vaatluste hulka, mille korral tunnus X võtab i-nda ning tunnus Y omandab j-nda väärtuse. Ridade sageduste summa

$$k_{i.} = \sum_{j=1}^{r} k_{ij}$$

ja veergude sageduste summa

$$k_{.j} = \sum_{i=1}^{l} k_{ij}.$$

Defineerime suuruse

$$\widetilde{k_{ij}} = \frac{k_{i.}k_{.j}}{n}.$$

Kui tunnused X ja Y on sõltumatud, siis

$$k_{ij} = \widetilde{k_{ij}}.$$

Seda arvestades koostame teststatistiku

$$H = \sum_{i=1}^{l}\sum_{j=1}^{r} \frac{(k_{ij} - \widetilde{k_{ij}})^2}{\widetilde{k_{ij}}} \sim \chi^2((l-1)(r-1)).$$

Olulisustõenäosus (*p-value*) leitakse χ^2-testide puhul järgmiselt. Olgu h statistiku H väärtus, mis on saadud mõõtmistulemuste põhjal. Siis

1) vasakpoolse hüpoteesi korral $p\text{-}value = P(H < h)$;

3) parempoolse hüpoteesi korral $p\text{-}value = P(H > h)$;

4) kahepoolse hüpoteesi korral $p\text{-}value = 2\min\{P(H \le h), P(H > h)\}$.

Näited erinevatest χ^2-testidest.

Näide 1.19. Mõõdeti 6 detaili läbimõõtu (mm). Saadi järgmised tulemused:

1.3	2.1	1.2	1.8	1.6	1.9

Kas võib nende tulemuste põhjal väita, et läbimõõdu dispersioon ületab 0.1 mm^2, kui eeldada, et detaili läbimõõt allub normaaljaotusele? Olulisuse nivooks olgu 0.05. Antud juhul on hüpoteeside paar järgmine:

$$\begin{cases} H_0 : \sigma^2 = 0.1, \\ H_1 : \sigma^2 > 0.1. \end{cases}$$

Leiame mõõtmistulemuste põhjal keskväärtuse ning dispersiooni nihketa hinnangud. Saame, et

$$\overline{x} = \frac{1.3 + 2.1 + 1.2 + 1.8 + 1.6 + 1.9}{6} = 1.65$$

ja

$$s^2 = \frac{1}{5} \sum_{i=1}^{6} (x_i - 1.65)^2 \approx 0.123.$$

Normaaljaotuse eeldusel statistik

$$H = \frac{s^2}{\sigma^2} \cdot 5 \sim \chi^2(5).$$

Antud mõõtmistulemuste põhjal saame selle statistiku väärtuseks $h \approx 6.15$. Sellele vastav olulisustõenäosus

$$p\text{-}value = P(H > 6.15) \approx 0.29 > 0.05.$$

Seega oleme sunnitud jääma nullhüpoteesi H_0 juurde. Tarkvara MS Excel abil saab antud juhul leida olulisustõenäosust funktsiooniga

CHISQ.DIST.RT(6.15;5).

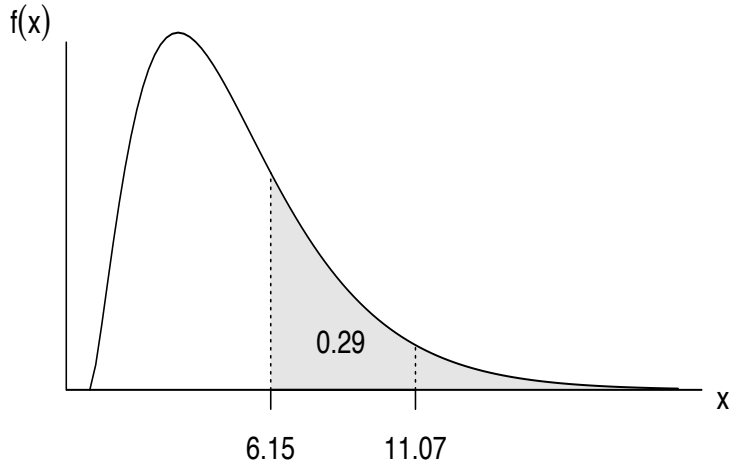
Väärtuse t_{crit_2} saab leida selle tarkvara funktsiooniga

CHISQ.INV.RT(0.05;5).

Antud juhul $t_{crit_2} \approx 11.07$. Oleme saanud piirkonnad

$$\mathcal{H}_t = (6.15; \infty), \quad \mathcal{H}_0 = [0; 11.07] \text{ ja } \mathcal{H}_1 = (11.07; \infty).$$

Allpool oleval joonisel on graafiliselt kujutatud antud näitel vastav piirkond $\mathcal{H}_t$.



Joonis 1.8. Piirkonnale $\mathcal{H}_t$ vastav olulisustõenäosus näite 1.19 korral

Näide 1.20. Käesolevas näites uuriti lambipirni eluiga T. Selleks leiti 150 pirni vastupidavusaeg tundides. Saadi järgmine empiiriline jaotus:

$[x_i; x_{i+1})$	n_i
$[0; 200)$	44
$[200; 400)$	37
$[400; 600)$	22
$[600; 800)$	15
$[800; 1000)$	11
$[1000; 1200)$	8
$[1200; 1400)$	7
$[1400; 1600]$	6

Formuleeriti hüpoteeside paar

$$\begin{cases} H_0 : T \sim \mathcal{E}(\nu), \\ H_1 : T \text{ allub muule jaotusele.} \end{cases}$$

Esmalt leiti parameetrile ν STP hinnang $\nu^* = \dfrac{1}{\overline{x}}$, kus

$$\overline{x} = \frac{1}{150} \sum_{i=1}^{8} 0.5(x_i + x_{i+1}) n_i \approx 490.7.$$

Seega $\nu^* \approx 0.002$. Seejärel leiti tõenäosus

$$\text{teor}_i = 150 p_i = 150(F(x_{i+1}) - F(x_i)) =$$

$$= 150\{1 - \exp(-\nu^* x_{i+1}) - (1 - \exp(-\nu^* x_i))\} =$$

$$= 150\{\exp(-\nu^* x_i) - \exp(-\nu^* x_{i+1})\}.$$

Saadud tulemustest koostati järgmine tabel:

$[x_i; x_{i+1})$	n_i	teor_i
$[0; 200)$	44	50.1
$[200; 400)$	37	33.4
$[400; 600)$	22	22.2
$[600; 800)$	15	14.8
$[800; 1000)$	11	9.9
$[1000; 1200)$	8	6.6
$[1200; 1400)$	7	4.4
$[1400; 1600]$	6	2.9

Selle tabeli põhjal leiti teststatistiku H väärtus

$$h = \sum_{i=1}^{8} \frac{(n_i - \text{teor}_i)^2}{\text{teor}_i} \approx 6.45.$$

Antud juhul allub teststatistik ligikaudu χ^2-jaotusele vabadusastmete arvuga 6. Järelikult olulisustõenäosus

$$p\text{-}value = P(H > 6.45) = \text{CHISQ.DIST.RT}(6.45; 6) \approx 0.37.$$

Seega tuleb jääda nullhüpoteesi juurde, mis väidab, et lambipirni eluiga allub eksponentjaotusele.

Näide 1.21. Uuriti mingi filmi meeldimise sõltuvust vanusest. Selleks defineeriti järgmised tunnused:

$$X = \begin{cases} 0, \text{ kui vanus on alla 18 eluaasta,} \\ 1, \text{ kui vanus on 18 ja 45 eluaasta vahel,} \\ 2, \text{ kui vanus on üle 45 eluaasta;} \end{cases}$$

$$Y = \begin{cases} 1, \text{ kui film meeldis,} \\ 0, \text{ kui filmi suhtuti neutraalselt,} \\ -1, \text{ kui film ei meeldinud.} \end{cases}$$

Püstitati järgmine hüpoteeside paar:

$$\begin{cases} H_0 : X, Y \text{ on sõltumatud,} \\ H_1 : X, Y \text{ ei ole sõltumatud.} \end{cases}$$

Kontrollimaks seda küsitleti 110 eri vanuses isikut. Tulemused esitati allpool toodud tabelina:

X \ Y	-1	0	1	$k_{i.}$
0	5	9	22	36
1	11	5	17	33
2	17	15	9	41
$k_{.j}$	33	29	48	$\Sigma = 110$

Selle tabeli põhjal saame teststatisiku H väärtuseks

$$h = \sum_{i=1}^{3} \sum_{j=1}^{3} = \frac{\left(k_{ij} - \frac{k_{i.}k_{.j}}{110}\right)^2}{\frac{k_{i.}k_{.j}}{110}} \approx 15.66.$$

Kuna nii tunnusel X kui ka tunnusel Y esineb 3 erinevat väärtust, siis allub teststatistik H hii-ruut-jaotusele vabadusastmete arvuga $4 = (3-1) \cdot (3-1)$. Seega olulisustõenäosus

$$p\text{-}value = P(H > 15.66) \approx 0.0035.$$

Seega võime kummutada sõltumatuse eelduse ning lugeda tõestatuks hüpoteesi H_1, mille kohaselt on vanus ja filmi meeldimine teineteisest sõltuvad.

Studenti t-test

Studenti t-testi näol on tegemist vast enim rakendust leidva statistilise testiga. Esitame definitsioonina jaotuse, mida nimetatakse Studenti t-jaotuseks.

Definitsioon 1.26. Olgu meil juhuslik suurus $Z \sim \mathcal{N}(0,1)$ ja juhuslik suurus $Y_n \sim \chi^2(n)$. Siis juhuslik suurus

$$T = \frac{Z}{\sqrt{\frac{Y_n}{n}}}$$

on t-jaotusega vabadusastmete arvuga n.

Juhusliku suuruse T allumist t-jaotusele vabadusastmete arvuga n tähistagem kui $T \sim t(n)$. Selle juhusliku suuruse α-kvantiili tähistus vabadusastmete arvu n korral olgu $t_{\alpha;n}$. Studenti t-jaotusele vastav keskväärtus ja dispersioon on järgmised:

$$E(T) = 0 \quad (n \geq 2)$$

ning

$$D(T) = \frac{n}{n-2} \quad (n \geq 3).$$

Studenti t-jaotuse tihedusfunktsiooniga võib huviline põhjalikumalt tutvuda õpiku [41] (lk 142–145) abiga. Lisaks t-jaotuse 2 tähtsat omadust:

1° Studenti t-jaotusele vastav tihedusfunktsioon f on paarisfunktsioon, see tähendab $f(-t) = f(t)$;

2° kui vabadusastmete arv $n \to \infty$, siis $t(n) \mapsto \mathcal{N}(0,1)$.

Omaduses 2° toodud asümptootika hakkab toimima piisavalt täpselt, kui $n \geq 120$. Seda asümptootikat saab uurida õpiku lõpus olevatest tabelitest lisa 1 ja lisa 2. Definitsioonidest 1.25 ja 1.26 ning lausest 1.2 järeldub järgmine tulemus.

Lause 1.5. Olgu juhuslik suurus $X_i \sim \mathcal{N}(\mu, \sigma)$. Siis

$$T = \frac{\overline{x} - \mu}{s}\sqrt{n} \sim t(n-1),$$

kus s tähistab juhusliku suuruse X_i standardhälbe nihketa hinnangut.

Lausel 1.5 põhinevad erinevad t-testid. Uurime neid lähemalt.

Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$. Oluline on eeldada, et $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., n$. Me saame normaaljaotuse parameetrile μ esitada hüpoteeside paari

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{cases}$$

Sellele hüpoteeside paarile vastab lauses 1.5 esitatud teststatistik, kus $\mu = \mu_0$.

Lauses 1.5 oleva statistiku abil saab konstrueerida normaaljaotusega juhusliku suuruse keskväärtusele α-usaldusintervalli järgmiselt:

$$I_\alpha = \left[\overline{x} - \frac{s}{\sqrt{n}} t_{\frac{1+\alpha}{2};n-1}; \overline{x} + \frac{s}{\sqrt{n}} t_{\frac{1+\alpha}{2};n-1}\right].$$

Seega oleme leidnud keskväärtuse α-usaldusintervalli konstrueerimiseks kaks meetodit:

1) usaldusintervalli konstrueerimine tsentraalse piirteoreemi rakendamise ehk normaalse aproksimatsiooni abil;

2) usaldusvahemiku leidmine Studenti t-jaotuse abil.

Alljärgnevas tabelis on toodud nende meetodite eelised ja puudused:

	Studenti t-test	Normaalne aproksimatsioon
Eelised	Valimi maht n võib olla väike	Jaotuste valik lai
Puudused	Eeldus, et $X \sim \mathcal{N}(\mu, \sigma)$	Valimi maht $n \geq 30$

Uurime Studenti t-testi rakendamist keskväärtuste võrdlemisel. Olgu meil valimid

$$\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal \text{ ja } \mathbf{Y} = (Y_1, Y_2, ..., Y_m)^\intercal.$$

Oluline on eeldada, et $X_i \sim \mathcal{N}(\mu_x, \sigma_x)$, $i = 1, 2, ..., n$ ning $Y_j \sim \mathcal{N}(\mu_y, \sigma_y)$, $j = 1, 2, ..., m$. Koostame hüpoteeside paari

$$\begin{cases} H_0 : \mu_x = \mu_y, \\ H_1 : \mu_x \neq \mu_y. \end{cases}$$

Selle kontrollimiseks kasutame t-jaotusele alluvat teststatistikut. Vaatleme kahte juhtu.

Test erinevate valimite puhul

Esmalt käsitleme juhtu, kus valimid $\mathbf{X}$ ja $\mathbf{Y}$ esindavad erinevaid objekte. See juht jaguneb omakorda kaheks.

1) Eeldame, et valimite $\mathbf{X}$ ja $\mathbf{Y}$ dispersioonid on võrdsed (ingl *equal variance*). Siis tõestatakse keskväärtuste erinevust teststatistikuga

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t(n+m-2),$$

kus s_x^2 ja s_y^2 tähistavad valimite $\mathbf{X}$ ja $\mathbf{Y}$ dispersioonide nihketa hinnanguid.

2) Olgu valimite $\mathbf{X}$ ja $\mathbf{Y}$ dispersioonid erinevad (ingl *unequal variance*). Siis keskväärtuste erinevust tõestav teststatistik

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}.$$

Vabadusastmete arv (df) leitakse järgmiselt:

$$df = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{s_x^4}{n^2}(n-1) + \frac{s_y^4}{m^2}(m-1)}.$$

Test kordusmõõtmiste puhul

Järgnevalt käsitleme juhtu, kus valimid $\mathbf{X}$ ja $\mathbf{Y}$ esindavad samu objekte erinevatel aegdel. Seega antud juhul $n = m$. Tegemist on n-ö paardunud

t-testiga (ingl *paired t-test*). Olukord vastab näiteks juhule, kus mõõdetakse samade patsientide vererõhku enne ja pärast mingi ravimi manustamist. Sellisel juhul leitakse valimi vahe $D_i = X_i - Y_i$, $i = 1, 2, ..., n$. Seejärel saadakse keskmise hinnang

$$\overline{d} = \frac{1}{n} \sum_{i=1}^{n} D_i$$

ning standardhälbe hinnang

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (D_i - \overline{d})^2}.$$

Vastav hüpoteeside paar on nüüd järgmine:

$$\begin{cases} H_0 : d = 0, \\ H_1 : d \neq 0. \end{cases}$$

Sellele paarile vastav teststatistik

$$T = \frac{\overline{d}}{s_d} \sqrt{n} \sim t(n-1).$$

Demonstreerime Studenti t-testi teostamist tarkvara MS Excel abil konkreetsete näidetega.

Näide 1.22. Uuriti kahe tööpingi tootlikkust (kg/tunnis). Saadi järgmised tulemused:

Tööpink A	15.1	14.8	14.7	14.9	15.3	15.0
Tööpink B	14.1	14.9	15.1	14.1	14.5	14.9

Püstitati küsimus, kas võib olulisuse nivool 0.05 ümber lükate väite, et nende kahe tööpingi tootlikkus on erinev. Kui eeldada, et mõlema tööpingi tootlikkus allub normaaljaotusele, siis võib anda vastuse t-testiga.

Studenti t-testi saab teostada suhteliselt hõlpsalt tarkvara MS Excel abil. Selle töövahendi (ingl *tool*) *Data Analysis* alt saab leida statistilised testid: *t-test*: *Paired Two Samples for Mean*; *t-test*: *Two-sample Assuming Equal Varainces* ja *t-test*: *Two-sample Assuming Unequal Varainces*.

Antud juhul tuleb valida kas *Two-sample Assuming Equal Varainces* või *Two-sample Assuming Unequal Varainces*. Tööpingi A puhul saadi keskväärtuseks ja dispersiooniks vastavalt

$$\overline{x}_A \approx 15.1 \text{ ja } s_A^2 \approx 0.35.$$

Tööpingi B puhul olid need karakteristikud järgmised:

$$\overline{x}_B \approx 14.6 \text{ ja } s_A^2 \approx 0.19.$$

Saadi järgmised olulisustõenäosused:

1) olulisustõenäosus *p-value* ≈ 0.1, kui valida test *Two-sample Assuming Unequal Varainces*;

2) olulisustõenäosus *p-value* ≈ 0.09, kui valida test *Two-sample Assuming Equal Varainces*.

Seega jäi hüpotees tööpinkide võrdsest tootlikkusest kummutamata.

Näide 1.23. Käesolevas näites uuriti ravimi mõju vererõhule. Mõõdeti viie patsiendi ülemist (ehk süstoolset) vererõhku enne ja pärast ravimi manustamist. Saadi järgmised tulemused:

Enne manustamist	171	168	185	189	191
Tund peale manustamist	139	140	165	161	159

Taheti kontrollida, kas antud ravim alandas oluliselt vererõhku. Antud juhul on tegemist samade indiviididega erinevatel aegadel ehk kordusmõõtmisega. Seega tuli valida test *Paired Two Samples for Mean*. Suurustele $\overline{d}$ ja s_d saadi järgmised väärtused:

$$\overline{d} = 28 \text{ ja } s_d \approx 4.9.$$

Antud teststatistiku väärtus

$$t \approx \frac{28}{4.9}\sqrt{5} \approx 12.8.$$

Lõpptulemuseks saadi, et vererõhu langus on oluline, sest olulisustõenäosus *p-value* < 0.0001.

Fisheri F-test

Uurime statistilise andmeanalüüsi ühte põhilist testi. Defineerime esmalt jaotuse, mida nimetatakse Fisheri F-jaotuseks ehk Fisheri jaotuseks.

Definitsioon 1.27. Olgu $Y_n \sim \chi^2(n)$ ja $Y_m \sim \chi^2(m)$. Siis juhuslik suurus

$$G = \frac{\frac{Y_n}{n}}{\frac{Y_m}{m}}$$

on F-jaotusega vabadusastmete arvudega n ja m.

Tähistagem Fisheri F-jaotusele allumist kui $G \sim F(n, m)$. Fisheri F-jaotusele vastava tihedusfunktsiooni graafik sarnaneb χ^2-jaotuse tihedusfunktsiooni graafikule. Kummagi jaotuse puhul ei ole graafik sümmeetriline. Üksikasjalikumat infot F-jaotusega juhusliku suuruse tihedusfunktsioonist võib huviline leida õpikust [41] (lk 145–147). Esitame siinkohal sellele jaotusele vastava keskväärtuse ja dispersiooni:

$$E(G) = \frac{m}{m-2}, \quad (m > 2)$$

ning

$$D(G) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}, \quad (m > 4).$$

Konstrueerime testi, mida nimetatakse Fisheri F-testiks. Olgu meil valimid $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ ja $\mathbf{Y} = (Y_1, Y_2, ..., Y_m)^\intercal$. Olgu $X_i \sim \mathcal{N}(\mu_x, \sigma_x)$, $i = 1, 2, ..., n$ ja $Y_j \sim \mathcal{N}(\mu_y, \sigma_y)$, $j = 1, 2, ..., m$. Esitame hüpoteeside paari

$$\begin{cases} H_0 : \sigma_x^2 = \sigma_y^2, \\ H_1 : \sigma_x^2 \neq \sigma_y^2. \end{cases}$$

Selle kontrollimiseks leiame dispersioonide nihketa hinnangud

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2$$

ja

$$s_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{y})^2.$$

Eelneva põhjal teame, et normaaljaotuse eeldusel statistik

$$H_x = \frac{s_x^2}{\sigma_x^2}(n-1) \sim \chi^2(n-1)$$

ning statistik

$$H_y = \frac{s_y^2}{\sigma_y^2}(m-1) \sim \chi^2(m-1).$$

Seega saame H_x ja H_y suhtest koostada teststatistiku

$$G = \frac{s_x^2}{s_y^2} \sim F(n-1, m-1), \tag{1.10}$$

mille parameetreid $n-1$ ja $m-1$ nimetatakse vabadusastmete arvudeks. Statistiku G abil testitakse normaaljaotusega juhuslike suuruste dispersioonide erinevust. Leiame selle statistiku kvantiilid

$$g_1 = g_{\frac{\beta}{2}}$$

ja

$$g_2 = g_{1-\frac{\beta}{2}}.$$

Kahepoolse hüpoteesi korral on otsustuskriteerium järgmine: kui leitud teststatistiku väärtus $g \notin [g_1; g_2]$, siis loeme tõestatuks hüpoteesi H_1, kui aga $g \in [g_1; g_2]$, siis oleme sunnitud jääma hüpoteesi H_0 juurde. Parempoolse hüpoteesi korral loeme H_1 tõestatuks juhul, kui $g > g_{1-\beta}$. Vasakpoolse hüpoteesi korral aga juhul, kui $g < g_\beta$.

Olulisustõenäosuse leidmine F-testi korral toimub analoogselt χ^2-testide juhuga.

Näide 1.24. Olulisuse nivool 0.05 taheti tõestada, et üks tööpink töötab täpsemalt kui teine. Selleks mõõdeti kahel tööpingil valmistatud detailide läbimõõtu (mm). Esimese pingi puhul saadi 8 detaili mõõtmistulemusteks

4.0	4.1	4.8	4.9	4.0	4.3	4.2	4.7

Teise pingi puhul saadi 6 detaili läbimõõtudeks

4.9	3.8	4.0	4.7	4.1	4.5

Eeldades, et mõlema tööpingi valmistatud detalide läbimõõdud alluvad normaaljaotusele, koostati hüpoteeside paar

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2, \\ H_1 : \sigma_1^2 \neq \sigma_2^2, \end{cases}$$

kus σ_1^2 ja σ_2^2 on vastavalt esimese ja teise tööpingi valmistatud detalide läbimõõdu dispersioonid. Mõõtmistulemuste põhjal saadi I pingi puhul läbimõõdu standardhälbe hinnanguks 0.37 ja II pingi puhul 0.43. Statistiku 1.10 väärtuseks osutus

$$G = \frac{0.43^2}{0.37^2} \approx 1.37.$$

Saadud väärtuse 1.37 põhjal tuleb teha otsustus. Antud juhul allub statistik G Fisheri jaotusele vabadusastmete arvudega 5 ja 7. Seega saame, et

$$g_1 = 0.15 \text{ ja } g_2 = 5.29.$$

Antud juhul kuulub teststatistiku G väärtus lõiku $[g_1; g_2]$. Seega oleme sunnitud jääma nullhüpoteesi H_0 juurde. Leiame antud mõõtmistulemustele vastava olulisustõenäosuse. Saame, et

$$p\text{-}value = 2\min\{P(G \leq 1.37); P(G > 1.37)\} \approx 0.68 > 0.05.$$

Tarkvara MS Excel abil saab leida kahepoolse F-testi korral olulisustõenäosust funktsiooniga F.TEST.

1.3.4. Statistiliste hüpoteeside kontrollimine tarkvara R abil

Tarkvara R näol on tegemist statistilise analüüsi teostamiseks koostatud programmeermiskeelega. See on vabavara, mille abil on võimalik lahendada laialdasel hulgal statistilisi probleeme (hüpoteeside kontroll, lineaarne

ja mittelineaarne regressioon, faktoranalüüs, mitteparameetriline statistika jne). Viimastel aastatel on R leidnud üha enam rakendust. Seda tarkvara saab alla laadida järgmiselt:

1) keskkonnas Windows lingilt
`https://cran.cnr.berkeley.edu/bin/windows/base/`;

2) keskkonnas Linux lingilt
`http://cran.rstudio.com/bin/linux/ubuntu`.

Järgnevalt näitame, kuidas viia läbi ülal kirjeldatud statistilisi hüpoteese tarkvara R keskkonnas.

Normaalne aproksimatsioon

Normaaljaotusele vastava jaotusfunktsiooni väärtuse leiab R-i käsk

`pnorm(z, mean = mu, sd = sigma)`,

kus `z` on argument ning `mean=mu` ja `sd=sigma` määravad vastavalt keskväärtuse μ ja standardhälbe σ. Kehtib seos

$$\texttt{pnorm(z)} = \Phi(z) + 0.5,$$

sest vaikimisi võetakse `mean=0` ja `sd=1`. Standardsele normaaljaotusele vastav tihedusfunktsioon leitakse käsuga

`dnorm(z)`

ning vastav α-kvantiil käsuga

`qnorm(alfa)`.

χ^2-testid

Realiseerimaks näites 1.19 toodud testi normaaljaotusega juhusliku suuruse dispersioonile tarkvaras R tuleb esmalt sellele installeerida pakett `TeachingDemos` käsuga

`install.packages("TeachingDemos")`.

Seejärel saab χ^2-testi läbi viia järgmiselt:

1) lugeda mõõtmistulemused sisse käsuga
`x=c(1.3,2.1,1.2,1.8,1.6,1.9)`;

2) teostada käsuga
`sigma.test(x,sigmasq=0.1,alternative="greater")`
soovitud χ^2-test.

Argument `alternative="greater"` tähendab, et tegemist on parempoolse hüpoteesiga.

Realiseerimaks sõltumatuse testi tarkvara R abil tuleb esmalt käsuga

`library(MASS)`

alla laadida vajalik tööpakett. Järgnevalt on vaja R-i keskkonda sisse lugeda tunnused X ja Y ning moodustada neist maatriks käsuga

`andmed=cbind(x,y)`.

Seejärel saab sõltumatuse testi teha järgmiselt:

```
tabel=table(andmed$x,andmed$y)
chisq.test(tabel).
```

Väljundiks saadakse olulisustõenäosus *p-value*.

Studenti t-testid

Demonstreerime Studenti t-testi realiseerimist näite 1.22 baasil.

Esmalt tuleb tarkvaras R mõõtmistulemused järgmiselt sisse lugeda:

```
x=c(15.1,14.8,14.7,14.9,15.3,15.0)
y=c(14.1,14.9,15.1,14.1,14.5,14.9)
```

Seejärel saab viia t-testi läbi käsuga

```
t.test(x,y,alternative="two.sided",mu=0,
paired=FALSE,var.equal=FALSE),
```

kus argumendiga `alternative` määratakse, kas tegemist on vasak- (väärtus `"less"`), parem- (`"greater"`) või kahepoolse (`"two.sided"`) hüpoteesiga, argument `paired` määrab, kas tegemist on kahe valimiga (`FALSE`) või kordusmõõtmistega (`TRUE`) ning argument `var.equal` näitab, kas eeldatakse valimite dispersioonide võrdsust (`TRUE`) või erinevust (`FALSE`).

Fisheri F-test

Näites 1.24 toodud statistilist testi saab teha R-i keskkonnas. Selleks tuleb sisestada käsk

`var.test(x,y,alternative="two.sided")`,

kus `x` ja `y` tähistavad sisse loetud diameetrite väärtusi. Argumendiga `alternative` saab nagu t-testigi puhul muuta kahepoolset hüpoteesi vasak- või parempoolseks.

1.4. Ülesanded

Ülesanne 1.1. Olgu meil valim $\mathbf{X} = (X_1, X_2, ...X_9)^\intercal$ üldkogumist, mille puhul keskväärtus on μ ja dispersioon σ^2. Olgu meil järgmised statistikud hindamaks parameetrit μ :

$$\overline{x}_1 = \frac{X_1 + X_2 + ... + X_9}{9}$$

ja

$$\overline{x}_2 = \frac{X_1 + 2X_5 + X_9}{2}.$$

Kas need hinnangud on nihketa? Mõjusad?

Ülesanne 1.2. Olgu meil parameetri θ hinnangud $\widehat{\theta}_1$ ja $\widehat{\theta}_2$. On teada, et $E(\widehat{\theta}_1) = \theta$ ning $E(\widehat{\theta}_2) = \frac{\theta}{2}$. On veel teada, et dispersioon $D(\widehat{\theta}_1) = 10$ ning dispersioon $D(\widehat{\theta}_2) = 4$. Milline hinnangutest on parem? Näpunäide: võtke headuse kriteeriumiks suurus MSE.

Ülesanne 1.3. Olgu meil juhuslikud suurused $X_1, X_2, ..., X_n$ sõltumatud ja normaaljaotusega keskväärtusega μ ning dispersiooniga σ^2. Olgu $\overline{x}$ määratud võrdusega

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Näidake, et juhuslik suurus

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$$

on nihketa hinnang dispersioonile σ^2.

Ülesanne 1.4. Kolm keemikut mõõtsid 2. järku keemilise reaktsiooni kiiruskonstanti k ($\mathrm{s}^{-1}\mathrm{M}^{-1}$). Mõõtmistulemusteks saadi vastavalt x_1, x_2 ning x_3. Tegemist on juhuslike suuruste $X_i \sim \mathcal{N}(k, 0.2)$, $i = 1, 2, 3$, realisatsioonidega. Õigeks tulemuseks loeti

$$\overline{x} = \frac{x_1 + x_2 + x_3}{3}.$$

Milline on juhusliku suuruse $\overline{X}$ jaotus?

Ülesanne 1.5. Kontrollimaks püssi headust lasevad n laskurit kuni esimese kümne silmani. Sihile jõudmiseks kulus järgmine hulk katseid: x_1, x_2, ..., x_n. Leidke suurima tõepära (STP) hinnang „kümne“ saamise tõenäosusele p. Eeldame laskude sõltumatust.

Ülesanne 1.6. Öeldakse, et juhuslik suurus X allub diskreetsele Pareto jaotusele parameetriga $a > 0$, kui tema tõenäosusfunktsioon

$$P(X = k) = \frac{a^k}{(1 + a)^{k+1}},$$

$k = 0, 1, ..., n, ...$ Olgu meil juhuslik suurus X, mis esindab diskreetse Pareto jaotusega valimit. Leidke selle realisatsiooni $\mathbf{x} = (x_1, x_2, ..., x_n)^\intercal$ põhjal STP hinnang parameetrile a.

Ülesanne 1.7. Konstrueerige normaalse aproksimatsiooni abil 0.95-usaldusvahemik eksponentjaotuse parameetrile. Olgu valimi maht $n = 36$.

Ülesanne 1.8. Olgu meil 6-elemendiline valim

$$\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)^\intercal$$

ning selle realisatsioon $\mathbf{x} = (4.5, 3.9, 4.1, 4.8, 5.1, 3.7)^\intercal$. Olgu selle valimi esindajaks juhuslik suurus $X \sim \mathcal{N}(\mu, \sigma^2)$. Leidke valimi realisatsiooni põhjal 0.95-usaldusintervall dispersioonile σ^2.

Ülesanne 1.9. Konstrueerige 0.95-usaldusintervall ülesandes 1.8 oleva juhusliku suuruse keskväärtusele μ.

Ülesanne 1.10. Olgu meil valim $(X_1, X_2, X_3, X_4, X_5)^\intercal$. Eeldame, et antud valimi elemendid alluvad normaaljaotusele. Valimi põhjal saadi standardhälbe nihketa hinnanguks 2.1 ühikut ning aritmeetiliseks keskmiseks 10 ühikut. Konstrueerige antud valimi põhjal 0.95-usaldusvahemik X_i keskväärtusele μ. Kas antud juhul jääb suhteline viga alla 10%?

Ülesanne 1.11. Mitu sõltumatut mõõtmist tuleb teha, et mõõtmistulemuste aritmeetiline keskmine ei erineks tõenäosusega 0.99 juhusliku suuruse keskväärtusest rohkem kui 5 ühiku võrra, kui standardhälve on 10 ühikut?

Ülesanne 1.12. Enne valimisi tehti küsitlus linnapea populaarsusest. Küsitlus toimus vormis „pooldan“ või „ei poolda“. Linnapea pooldajate osakaaluks saadi 0.55. Leidke minimaalne küsitletute hulk n, mille puhul pooldajate osakaalu 0.95-usaldusintervalli alumine piir on üle 0.5.

Ülesanne 1.13. Olgu meie eesmärgiks leida 100 küsitletu põhjal 0.95-usaldusintervall osakaalu hinnangule $\widehat{p}$. Millise osakaalu väärtuse korral on see intervall kõige laiem ehk millise hinnangu korral on suhteline viga suurim?

Ülesanne 1.14. Allugu patarei eluiga aastates eksponentjaotusele, mille parameetriks on m. Uuriti viit patareid ning saadi vastupidavusajad (aastates) $\{0.9, 1.2, 1.6, 0.8, 1.1\}$. Leidke STP hinnang parameetrile m. Kui suur on leitud hinnangu põhjal tõenäosus, et patarei elab üle aasta?

Ülesanne 1.15. Kvaliteedikontrolli insener mõõtis 25 klaaspudeli seina paksust. Antud valimi põhjal sai ta keskväärtuse hinnanguks $\overline{x} = 4.058$ mm ja standardhälbe hinnanguks $s = 0.081$ mm.

1) Eeldame, et kvaliteedikontrolli insener pidi tõestama, et klaaspudeli seina paksus on üle 4 mm. Formuleerige antud tõestusele vastav statistiline test. Milline oleks järeldus olulisuse nivool $\beta = 0.05$? Milline on olulisustõenäosus?

2) Leidke antud valimi põhjal pudeli seina paksusele ning selle standardhälbele 0.95-usaldusintervallid.

Ülesanne 1.16. Kaugusmõõturiga mõõdeti 10 korda teatud vahemaa meetrites. Saadi järgmised tulemused:

202	195	207	203	194	209	198	208	204	192

Leidke mõõdetava vahemaa keskväärtuse ja dispersiooni nihketa hinnangud ning nende 0.95-usaldusintervallid. Eeldades, et kauguse mõõturil puudub süstemaatiline viga ning kaugused alluvad normaaljaotusele.

Ülesanne 1.17. Allugu detaili läbimõõt (mm) X_i, $i = 1, 2, ..., n$, normaaljaotusele $\mathcal{N}(\mu, 0.3)$. Parameetri μ kohta püstitati hüpoteesid

$$\begin{cases} H_0 : \mu = 10, \\ H_1 : \mu \neq 10. \end{cases}$$

Neid hüpoteese asusid kontrollima insener A ning insener B.

Insener A leidis valimi mahu $n = 8$ põhjal detaili läbimõõdu aritmeetilise keskmise $\overline{x}$. Ta püstitas järgmise otsustuskriteeriumi: kui $\overline{x} \in [9.9; 10.1]$, siis tuleb jääda H_0. Vastasel juhul lugeda tõestatuks H_1.

Insener B leidis aritmeetilise keskmise $\overline{x}$ valimi mahu $n = 12$ põhjal. Tema otsustuskriteerium oli järgmine: kui $\overline{x} \in [9.8; 10.2]$, siis tuleb jääda H_0 juurde. Vastasel juhul lugeda tõestatuks H_1.

Leidke mõlema inseneri otsustuskriteeriumile vastav võimsusfunktsioon $h(\mu)$. Milline on mõlema inseneri puhul $h(10.2)$? Kumma inseneri otsustuskriteeriumile vastava statistilise testi võimsus oli suurem?

Ülesanne 1.18. Ühes linnas kontrolliti 100 päeva jooksul toimunud veeavariide arvu. Saadi järgmised tulemused:

Avariide arv	Sagedus
0	8
1	28
2	31
3	18
4	9
5	6

Mis jaotusega on teie meelest veeavariide arv? Leidke suurima tõepära (STP) hinnang selle jaotuse parameetri(te)le. Konstrueerige teststatistik kontrollimiseks, kas empiiriline jaotus vastab teie pakutud teoreetilisele jaotusele. Olulisuse nivooks võtke 0.05.

Ülesanne 1.19. Eeldades, et ülesandes 1.18 allub veeavariide arv X Poissoni jaotusele, püstitati järgmine hüpoteeside paar parameetrile λ :

$$\begin{cases} H_0 : \lambda = 1.5, \\ H_1 : \lambda > 1.5. \end{cases}$$

Kas võib olulisuse nivool 0.05 kummutada nullhüpoteesi H_0? Aluseks võtta ülesandes 1.18 esitatud andmetabel.

Ülesanne 1.20. Mõõdeti rahvusgruppide A ja B intelligentsuse koefitsiendi IQ-näitajat. Igast rahvusgrupist valiti testimiseks 10 indiviidi. Saadi järgmised tulemused:

Rahvuse A IQ	Rahvuse B IQ
101	98
104	102
79	112
98	105
88	104
111	92
112	91
99	101
96	92
75	109

Leidke vastused järgmistele küsimustele.

1) Millised oleksid hüpoteeside paarid testimaks IQ-näitajate erinevust erinevate rahvusgruppide vahel?

2) Millist testi kasutate nullhüpoteesi kummutamiseks?

3) Mida me eeldame valitud testi kasutamisel?

4) Kas saate olulisuse nivool 0.1 nullhüpoteesi ümber lükata?

Ülesanne 1.21. Kontrollimaks päevaste askelduste mõju mõõdeti 8 inimese pikkust (cm) kell 9.00 ja kell 21.00. Saadi järgmised tulemused:

Kell 9.00	172	169	180	181	160	163	165	186
Kell 21.00	172	167	177	180	159	161	165	184

Kas nende tulemuste põhjal võib väita, et inimese pikkus kahanes päeva jooksul oluliselt? Olulisuse nivooks võtta 0.05.

Ülesanne 1.22. Olgu meil järgmine hüpoteeside paar keskväärtusele μ :

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu > \mu_0. \end{cases}$$

Leidke olulisustõenäosused järgmistele Z-statistiku väärtustele:

1) $z = 2.35$;

2) $z = 1.53$;

3) $z = 0.98$.

Ülesanne 1.23. Olgu uuritav suurus $X \sim \mathcal{N}(\mu, \sigma)$. Dispersioonile σ^2 esitati hüpoteeside paar

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2, \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{cases}$$

Hüpoteesi kontrollimiseks koostati valim mahuga 10. Leidke olulisustõenäosused statistiku H järgmistele väärtustele:

1) $h = 15.1$;

2) $h = 22.4$;

3) $h = 28.9$.

Ülesanne 1.24. Uuriti Mari ja Jüri jaanuarikuist päevast rahakulu. Jüri päevase rahakulu standardhälve oli 15 eurot, Maril aga oli päevase rahakulu standardhälbeks 10 eurot. Kas võib olulisuse nivool 0.05 väita, et Mari kulutas oma raha stabiilsemalt kui Jüri? Stabiilisuse mõõduks võtta standardhälve.

Ülesanne 1.25. Tekstiilikiude tootev firma uurib uue kanga lõnga vastupidavust, mõõtes pikenemise jõudu sellele lõngale rakendatava massi abil. Olgu selle massi standardhälve 0.3 kg. Firma testib 5 kanga näidise põhjal järgmist hüpoteesi massile μ (kg):

$$\begin{cases} H_0 : \mu = 14, \\ H_1 : \mu < 14. \end{cases}$$

1) Leidke olulisustõenäosus, kui antud valimi põhjal saadud aritmeetiline keskmine $\overline{x} = 13.7$ kg.

2) Leidke II liiki vea tegemise tõenäosus γ_2, kui tegelik pikenemise jõud on 13.5 kg ning olulisuse nivoo $\beta = 0.05$.

3) Milline on punktis 2 toodud tingimuste korral testi statistiline võimsus?

Ülesanne 1.26. Täringut visati 40 korda. Kuus silma saadi 8 korral.

1) Kas võib selle põhjal olulisuse nivool 0.05 väita, et täring ei ole sümmeetriline?

2) Mitu viset minimaalselt tuleks teha, et kuue silma sagedusel $\frac{1}{5}$ saaks antud visete põhjal ümber lükata täringu sümmeetria eelduse, kui olulisuse nivooks võtta 0.05?

Ülesanne 1.27. Laserkaugusmõõtja standardhälbeks saadi 10 mõõtmise põhjal 12 mm. Pärast remonti katsetati laserkaugusmõõtja täpsust uuesti. Nüüd saadi 12 mõõtmise põhjal standardhälbeks 8 mm. Kas võib olulisuse nivool 0.05 väita, et pärast remonti laserkaugusemõõtja täpsus oluliselt paranes?

Ülesanne 1.28. Sportlasel A ja sportlasel B olid hooaja 7 parema võistluse kettaheite tulemused järgmised:

Sportlane A	68.24	66.15	64.13	63.98	62.12	61.98	61.22
Sportlane B	65.66	65.12	64.02	63.28	63.12	63.04	62.55

Kas võib olulisuse nivool 0.05 ümber lükata väite, et mõlema sportlase stabiilsus oli hooajal võrdne? Eeldatakse, et sportlaste tulemused allusid selle hooaja jooksul normaaljaotusele.

2. peatükk
Andmeanalüüs

Eelmine peatükk käsitles matemaatilise statistika aluseid. Käesolevas peatükis rakendame matemaatilise statistika alaseid teadmised reaalsetele andmetele. Eesmärk on anda ülevaade statistilise andmeanalüüsi erinevatest meetoditest. Uurime neist lähemalt kolme peamist:

1) regressioonanalüüsi,

2) dispersioonanalüüsi,

3) faktoranalüüsi.

Enne statistilise andmeanalüüsi juurde asumist tuleb viia andmestik selleks sobivale kujule, milleks on üks maatriks. Kirjeldame lähemalt, milline on tema dimensioon ning millised on selle maatriksi elemendid.

2.1. Objekt-tunnus maatriks

Statistiline andmeanalüüs saab alguse struktuurist, mida nimetatakse objekt-tunnus maatriksiks. Tegemist on $n \times (k+1)$-maatriksiga, mille ridadeks on uurimisobjektid ning veergudeks tunnused. Selle elementideks on meid huvitavate objektide erinevate tunnuste mõõtmistulemused. Objekt-tunnus maatriksi struktuur on esitatud alljärgneva tabelina:

$\mathbf{Y}$	$\mathbf{X}_1$	$\cdots$	$\mathbf{X}_j$	$\cdots$	$\mathbf{X}_k$
y_1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1k}
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
y_i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ik}
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
y_n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{nk}

Tunnused võib jagada laias laastus kaheks: pidevad ning diskreetsed tunnused. Diskreetseteks tunnusteks võivad olla arvtunnused ehk järjestatavad tunnused või koodtunnused ehk mittejärjestatavad tunnused.

Objekt-tunnus maatriksi esimene veerg kujutab endast tavaliselt uuritava tunnuse vektorit $\mathbf{Y} = (Y_1, Y_2, ..., Y_i, ..., Y_n)^\intercal$, ülejäänud k veergu aga faktorite vektoreid

$$\mathbf{X}_j = (X_{1j}, X_{2j}, ..., X_{ij}, ..., X_{nj})^\intercal, \quad j = 1, 2, ..., k.$$

Tunnuseid $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k$ nimetatakse faktortunnusteks. Maatriksit

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{k1} & \cdots & X_{kk} \end{pmatrix}$$

nimetatakse plaanimaatriksiks. Meid huvitab, milline funktsionaalne seos võiks esineda uuritava tunnuse $\mathbf{Y}$ ning faktortunnuste $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k$ vahel. See tähendab sellise funktsiooni h leidmist, mille korral

$$Y_i \approx h(X_{i1}, X_{i2}, ..., X_{ik}), \quad i = 1, 2, ..., n.$$

Näide 2.1. Olgu meil objektideks 8 detaili. Meie eesmärk on uurida selle detaili massi (grammides) $\mathbf{Y}$ seoseid detaili läbimõõdust (mm) $\mathbf{X}_1$ ning vanuseklassist (0 – noor, 1 – vana) $\mathbf{X}_2$. Pärast detaili vanuse hindamist, läbimõõdu mõõtmist ja kaalumist võib saadud tulemused esitada järgmise objekt-tunnus maatriksina:

$\mathbf{Y}$	$\mathbf{X}_1$	$\mathbf{X}_2$
0.38	5.1	0
0.44	5.5	1
0.39	4.9	1
0.58	5.6	1
0.33	4.9	0
0.39	5.0	0
0.42	5.1	0
0.41	4.9	0

Meie eesmärk on leida selle tabeli põhjal parim seos

$$\mathbf{Y} = h(\mathbf{X}_{i1}, \mathbf{X}_{i2}) \quad i = 1, 2, ..., 8.$$

Selle parima seose leidmisega tegeleb regressioonanalüüs. Selles rakendatakse tõenäosusteooria ja matemaatilise statistika tulemusi. Regressioonanalüüsi võib jagada üldiste lineaarsete mudelite teooriaks ja üldistatud lineaarsete mudelite teooriaks.

Teeme järgnevalt tutvust lineaarsete mudelite struktuuriga, nende koostamise ning diagnoosimisega.

2.2. Üldise lineaarse mudeli struktuur

Üldise lineaarse mudeli puhul eeldame, et uuritav tunnus $Y \sim \mathcal{N}(\mu, \sigma)$. Juhuslik suurus Y on teatavasti valimi $\mathbf{Y} = (Y_1, Y_2, ..., Y_i, ..., Y_n)^\intercal$ suvalise elemendiga identse jaotusega koopia. Üldine lineaarne mudel avaldub kujul

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i, \tag{2.1}$$

kus mudeli jääk ehk prognoosijääk (ingl *residual*) $\epsilon_i \sim \mathcal{N}(0, \sigma)$. Mudeli (2.1) maatrikskujuline esitus on järgmine:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i1} & \cdots & X_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Lühidalt kirja panduna saame, et

$$\mathbf{Y} = \mathbf{Z}\beta + \epsilon,$$

kus $\mathbf{Y} : n \times 1$ on uuritava tunnuse valimi vektor, $\mathbf{Z} : n \times (k+1)$ tähistab plaanimaatriksit, millele on lisatud ühtede veerg, $\beta : (k+1) \times 1$ on parameetrite vektor ning $\epsilon : n \times 1$ tähistab mudeli prognoosijääkide vektorit. Nimetagem maatriksit $\mathbf{Z}$ lineaarse mudeli maatriksiks ehk lihtsalt mudeli maatriksiks. Leiame uuritava suuruse Y keskväärtuse. Eeldusel, et $E(\epsilon_i) = 0$, saame, et

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

ehk vektorkujul

$$\mu_i = \mathbf{z}_i \beta, \tag{2.2}$$

kus $\mu_i = E(Y_i)$ ning $\mathbf{z}_i : 1 \times k+1$ tähistab maatriksi $\mathbf{Z}$ reavektorit. Erinevalt mudelist (2.1) on mudelis (2.2) juhuslike suuruste asemel nende mõõdetud väärtused ehk juhuslike suuruste mingid realisatsioonid.

2.2.1. Mudeli parameetrite leidmine vähimruutude meetodil

Püstitame küsimuse, kuidas leida parameetrite vektorit β. Selle vektori leidmiseks kasutame vähimruutude (ingl *least squares*) meetodit. Olgu meil uuritava suuruse mõõdetud väärtused $y_1, y_2, ..., y_n$ ja mudeli (2.2) abil saadud väärtused $\mathbf{z}_1\beta, \mathbf{z}_2\beta, ..., \mathbf{z}_n\beta$. Meie eesmärgiks on minimiseerida funktsioon

$$Q = \sum_{i=1}^{n} (y_i - \mathbf{z}_i \beta)^2.$$

See tähendab, et peame leidma parameetrite $\beta_0, \beta_1, ..., \beta_k$ sellised väärtused, mille korral mudeli ja empiiriliste väärtuste vahe ruut oleks minimaalne. Sellest ka nimetus vähimruutude meetod. Eeldades funktsiooni Q nõgusust, tuleb meil lahendada võrrandisüsteem

$$\frac{\partial Q}{\partial \beta_j} = 0, \quad j = 0, 1, ..., k. \tag{2.3}$$

Uurime esmalt süsteemi (2.3) lahendit juhul, kui mudelis on üksnes parameeter β_0 ehk n-ö nullmudeli puhul. Siis saame, et

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}(y_i-\beta_0)^2=0,$$

millest on lihtne järeldada, et

$$\beta_0=\frac{1}{n}\sum_{i=1}^{n}y_i.$$

Seega on parameetri β_0 parimaks hinnanguks vähimruutude mõttes mõõtmistulemuste aritmeetiline keskmine.

Vaatleme nüüd lihtsat lineaarset regressiooni, mille korral $k=1$ ehk juhtu, millal mudelis (2.2) on 1 faktortunnus. Olgu selle faktortunnuse mõõtmistulemused $x_1, x_2, ..., x_n$. Siis avaldub süsteem (2.3) kujul

$$\begin{cases}\dfrac{\partial}{\partial\beta_0}\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2=0\\ \dfrac{\partial}{\partial\beta_1}\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2=0.\end{cases}$$

Leides mõlemad osatuletised ning tehes mõningad teisendused, saame järgmise lineaarse võrrandisüsteemi parameetrite β_0 ja β_1 leidmiseks

$$\begin{cases}n\beta_0+\beta_1\displaystyle\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i\\ \beta_0\displaystyle\sum_{i=1}^{n}x_i+\beta_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i.\end{cases}$$

Lahendades selle süsteemi (näiteks Crameri determinantide meetodil), saame, et

$$\beta_0=\frac{\displaystyle\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{\displaystyle n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

ning

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Olgu $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ ning $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$. Siis

$$\beta_0 = \frac{\overline{y}\ \overline{x^2} - \overline{x}\ \overline{xy}}{\overline{x^2} - (\overline{x})^2}$$

ja

$$\beta_1 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - (\overline{x})^2}.$$

Siinkohal tuleb meelde tuletada Pearsoni korrelatsioonikordaja definitsiooni ning selle kordaja omadusi.

Definitsioon 2.1. Suurust

$$\mathrm{corr}(X_j, Y) = \frac{\mathrm{cov}(X_j, Y)}{\sigma_{X_j}\sigma_Y}$$

nimetatakse Pearsoni korrelatsioonikordajaks.

Selle suuruse lugejat

$$\mathrm{cov}(X_j, Y) = E((X_j - E(X_j))(Y - E(Y)))$$

nimetatakse juhuslike suuruste X ja Y vaheliseks kovariatsiooniks.

Definitsiooni 2.1 abil saab tõestada, et $-1 \leq \mathrm{corr}(X, Y) \leq 1$. Tõestuse võib huviline leida õpikutest [34] (lk 107) või [41] (lk 118).

Anname kovariatsioonile ja korrelatsioonile geomeetrilise interpreteeringu. Pearsoni korrelatsioonikordajat saab interpreteerida kui kahe vektori vahelise nurga kosiinust. Kovariatsiooni interpreteeringuks aga sobib kahe vektori skalaarkorrutis. Standardhälbeid saab aga interpreteerida kui

vektorite pikkuseid. Tõepoolest, eukleidilises ruumis saame defineerida skalaarkorrutise vektorite $\overrightarrow{x}$ ja $\overrightarrow{y}$ vahel kui

$$\langle \overrightarrow{x}, \overrightarrow{y} \rangle = |\overrightarrow{x}||\overrightarrow{y}|\cos(\alpha),$$

kus α tähistab vektorite $\overrightarrow{x}$ ja $\overrightarrow{y}$ vahelist nurka.

Pearsoni korrelatsioonikordajat saab üldistada juhuslikule vektorile $\mathbf{X}$ mitmese korrelatsioonikordajana, mis avaldub kui

$$\mathbf{r} = \sqrt{\mathrm{corr}^\intercal(\mathbf{X}, Y)\mathbf{R}^{-1}\mathrm{corr}(\mathbf{X}, Y)},$$

kus

$$\mathrm{corr}(\mathbf{X}, Y) = \begin{pmatrix} \mathrm{corr}(X_1, Y) \\ \mathrm{corr}(X_2, Y) \\ \vdots \\ \mathrm{corr}(X_k, Y) \end{pmatrix}$$

ja maatriksi $\mathbf{R} : k \times k$ element

$$r_{ij} = \mathrm{corr}(X_i, X_j).$$

Maatriksit $\mathbf{R}$ nimetatakse tunnuste $X_1, X_2, ..., X_k$ vaheliseks korrelatsioonimaatriksiks.

Lause 2.1. Korrelatsioonimaatriksil on järgmised omadused:

1° korrelatsioonimaatriks $\mathbf{R}$ on sümmeetriline, s.t $\mathbf{R}^\intercal = \mathbf{R}$;

2° korrelatsioonimaatriksi $\mathbf{R}$ iga element $-1 \leq r_{ij} \leq 1$;

3° korrelatsioonimaatriks $\mathbf{R}$ on positiivselt (pool)määratud, s.t iga vektori $\mathbf{x}$ korral $\mathbf{x}^\intercal\mathbf{R}\mathbf{x} \geq 0$.

Avaldame parameetri β_1 vähimruutude hinnangu Pearsoni korrelatsioonikordaja kaudu. Olgu $\overline{s^2}_x$ ja $\overline{s^2}_y$ juhuslike suuruste X ja Y dispersioonide nihkega hinnangud. Statistik $\overline{s}_{xy} = \overline{xy} - \overline{x}\,\overline{y}$ kujutab endast juhuslike suuruste X ja Y vahelise kovariatsiooni nihketa hinnangut. Siis saame Pearsoni korrelatsioonikordaja hinnanguks

$$r_{xy} = \frac{\overline{s}_{xy}}{\sqrt{\overline{s^2}_x}\sqrt{\overline{s^2}_y}}.$$

Kuna $\overline{s^2}_x = \overline{x^2} - \overline{x}^2$ ja $\overline{s^2}_y = \overline{y^2} - \overline{y}^2$, siis parameeter

$$\beta_1 = \frac{\overline{s}_{xy}}{\overline{s^2}_x} = \frac{r_{xy}\sqrt{\overline{s^2}_x}\sqrt{\overline{s^2}_y}}{\overline{s^2}_x} = r_{xy}\sqrt{\frac{\overline{s^2}_y}{\overline{s^2}_x}}.$$

Saime, et

$$\beta_1 = r_{xy}\frac{\overline{s}_y}{\overline{s}_x}. \tag{2.4}$$

Järelikult võrdub kordaja β_1 parim hinnang vähimruutude mõttes tunnuste Y ja X vahelise lineaarse korrelatsioonikordaja hinnangu ja nende tunnuste standardhälvete (nihkega) hinnangute suhte korrutisega.

Uurime nüüd vähimruutude meetodit üldjuhul. Olgu meil mõõtmistulemuste vektor $\mathbf{y}$. Leiame funktsiooni Q tuletise parameetrite vektori β järgi. Saame, et

$$\frac{\partial Q}{\partial \beta} = \frac{\partial}{\partial \beta}(\mathbf{y} - \mathbf{Z}\beta)^\intercal(\mathbf{y} - \mathbf{Z}\beta) = -2\mathbf{Z}^\intercal(\mathbf{y} - \mathbf{Z}\beta) = \mathbf{0}_{k+1},$$

kus $\mathbf{0}_{k+1}$ tähistab $k+1$-komponendilist nullvektorit. Pärast mõningaid maatriksteisendusi saame parameetrite vektori β parimaks vähimruutude mõttes parimaks hinnanguks

$$\widehat{\beta} = (\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{y}. \tag{2.5}$$

2.2.2. Üldise lineaarse mudeli diagnostika

Viime läbi üldise lineaarse mudeli diagnostika. Vaatleme selle diagnoosimise 3 põhietappi.

Mudeli olulisuse kontroll

Kui mudel ei ole oluline, siis pole teda mõtet edasi uurida. Sõnastame hüpoteeside paari

$$\begin{cases} H_0 : \text{mudel ei ole oluline,} \\ H_1 : \text{mudel on oluline.} \end{cases}$$

Olgu empiirilise uuritava suuruse Y realisatsioon

$$y_1, y_2, ..., y_n$$

ning olgu teoreetilise uuritava suuruse $\widehat{Y}$ (ehk mudeliga (2.2) saadud) realisatsioon

$$\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n.$$

Jagame uuritava tunnuse Y koguhajuvuse S_y kaheks: regressioonist tingitud hajuvuseks S_{reg} ning mudeli jäägist tingitud hajuvuseks S_{res}. Saame, et

$$S_y = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 =$$

$$= \sum_{i=1}^{n}\{(y_i - \widehat{y}_i)^2 + 2(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + (\widehat{y}_i - \overline{y})^2\} =$$

$$= \sum_{i=1}^{n}((y_i - \widehat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2.$$

Eeldades, et $E(\epsilon_i) = 0$, saame, et

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = 0.$$

Olgu

$$S_{reg} = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

regressioonihajuvus ja

$$S_{res} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

mudeli jäägi hajuvus. Seega saime, et

$$S_y = S_{reg} + S_{res}.$$

Meid huvitab kumma hajuvuse, kas hajuvuse S_{reg} või hajuvuse S_{res}, osakaal on uuritava tunnuse koguhajuvuses suurem. Mudeli olulisuse kontrollimiseks koostatakse seega teststatistik, mis sisaldab endas S_{reg} ja S_{res} suhet. Küsime: mis jaotusega on statistik? Lause 1.2 põhjal saame, et

$$\frac{S_y}{\sigma^2} \sim \chi^2(n-1); \quad \frac{S_{reg}}{\sigma^2} \sim \chi^2(k); \quad \frac{S_{res}}{\sigma^2} \sim \chi^2(n-k-1).$$

Seda arvestades saame seose (1.10) abil soovitud statistikuks

$$G = \frac{S_{reg}(n-k-1)}{S_{res}k} \sim F(k, n-k-1).$$

Kui mudelis esineb 1 faktortunnus, siis

$$G \sim F(1, n-2).$$

Mudel loetakse oluliseks olulisuse nivool β, kui andmete põhjal leitud teststatistiku väärtus $g > g_{1-\beta}$. Kui mudel ei osutunud oluliseks, siis lõpetatakse selle kui halva mudeli analüüs. Kui mudel osutus oluliseks, siis kerkib üles uus küsimus.

Mudeli kordajate olulisuse kontroll

Selles diagnoosimise etapis eraldatakse olulised parameetrid mitteolulistest. Mudeli parameetrite $\beta_0, \beta_1, ..., \beta_k$ olulisust kontrollitakse hüpoteesside paariga

$$\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0, \end{cases}$$

kus $j = 0, 1, ..., k$. Osutub, et seda hüpoteeside paari kontrollitakse t-jaotuse abil. Olgu $\widehat{\beta}_j$ seose (2.5) põhjal leitud hinnang parameetrile β_j. Olgu meil j-nda faktortunnuse hajuvus

$$S_{x_j} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2,$$

kus

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij},$$

$j = 1, 2, ..., k$. Siis saame teststatistikuks

$$T_{\beta_j} = \frac{\widehat{\beta}_j}{\sqrt{\frac{S_{res}}{(n-2)S_{x_j}}}} \sim t(n-2), \quad j = 1, 2, ...k.$$

Kui $j = 0$, siis

$$T_{\beta_j} = \frac{\widehat{\beta}_j}{\sqrt{\frac{S_{res}}{(n-2)}}} \sim t(n-2).$$

Statistik T_{β_0} testib mudeli vabaliikme (ingl *intercept*) olulisust.

Mudelisse jäetakse üksnes need parameetrid, mille puhul leidis tõestamist hüpotees H_1. Ülejäänud parameetrite kohta öeldakse, et need ei erine oluliselt nullist ehk nendele vastavate faktortunnuste mõju võib olla nii negatiivne kui ka positiivne.

Kolmandaks uurime, kuivõrd kirjeldab mudel tegelikkust.

Mudeli kirjeldusvõime leidmine

Meid huvitab, kui suurt osa uuritava suuruse muutlikkusest kirjeldab lineaarne mudel. Mudeli kirjeldusvõimet iseloomustab suurus

$$R^2 = \frac{S_{reg}}{S_y} \in [0; 1].$$

Uuritava suuruse Y hajuvust kirjeldab $R^2 \cdot 100\%$ mudel (2.1) ja $(1 - R^2) \cdot 100\%$ selle mudeli jääk ϵ_i. Kordajat R^2 nimetatakse R-ruut (ingl *RSquared*) determinatsioonikordajaks. Formuleerime ja tõestame alljärgnevalt selle kordaja ühe omaduse.

Lause 2.2. Kehtib seos

$$R^2 = r^2_{y\widehat{y}},$$

kus $r^2_{y\widehat{y}}$ tähistab uuritava suuruse mõõdetud väärtuse Y ja lineaarse mudeliga saadud väärtuse $\widehat{Y}$ vahelise Pearsoni korrelatsioonikordaja hinnangut.

Tõestus. Ühelt poolt saame, et

$$R^2 = \frac{S_{reg}}{S_y} = 1 - \frac{S_{res}}{S_y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i^2 - 2y_i\widehat{y}_i + \widehat{y}_i^2)}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Kuna $\sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0$, siis $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$. Alati on võimalik koordinaate nihutada nii, et $\overline{y} = \overline{\widehat{y}}$. Selline nihutamine ei muuda dispersioone ning lineaarset korrelatsioonikordajat. Nii tehes saame, et

$$R^2 = 1 - \frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

Teisalt aga $\overline{y} = \overline{\widehat{y}}$ korral

$$r_{y\widehat{y}}^2 = \frac{\left(\sum_{i=1}^{n} y_i \widehat{y}_i\right)^2}{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} \widehat{y}_i^2} = \frac{\left(\sum_{i=1}^{n} \widehat{y}_i^2\right)^2}{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} \widehat{y}_i^2} = \frac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

Seega tõepoolest

$$R^2 = r_{y\widehat{y}}^2.$$

□

Lausest 2.2 järeldub, et R-ruut determinatsioonikordaja iseloomustab uuritava suuruse empiiriliste väärtuste ja mudeliga saadud väärtuste vahelise lineaarse seose tugevust.

Järeldus 2.1. Kui $R^2 = 1$, siis $|\operatorname{corr}(Y, \widehat{Y})| = 1$. Kui aga $R^2 = 0$, siis $\operatorname{corr}(Y, \widehat{Y}) = 0$.

Sageli kasutatakse R-ruut determinatsioonikordaja asemel korrigeeritud determinatsioonikordajat (ingl *adjusted R-Square*)

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}.$$

Suuruse R_{adj}^2 näol on tegemist R-ruut determinatsioonikordaja nihketa hinnanguga.

Näide 2.2. Demonstreerime üldise lineaarse mudeli koostamist ning diagnostikat Peipsi järve andmete põhjal. Bioloogia andmed kuuluvad Eesti Maaülikooli limnoloogiakeskusele, keemia andmed osaliselt ka Keskkonnaagentuurile. Need on kogutud teadusprojektide ja riikliku seire raames. Andmestik koosneb 156 mõõtmisest järve eri punktides. Uuriti Peipsi järve sinivetikate kontsentratsiooni (mg/liiter) sõltuvust järgmistest faktoritest: bakteroplankton (BAC), klorofüll (CHL), mineraalne fosfor (DIP), orgaaniline fosfor (DOP), räni (SI), mineraalne lämmastik (DIN) ja orgaaniline lämmastik (DON). Modelleerimiseks kasutati tarkvara MS Excel töövahendi *Data Analysis* protseduuri *Regression*. Koostati järgmine mudel:

$$\mathrm{CY} = \beta_0 + \beta_1 \cdot \mathrm{BAC} + \beta_2 \cdot \mathrm{CHL} + \beta_3 \cdot \mathrm{DIP} + \beta_4 \cdot \mathrm{DOP} + \beta_5 \cdot \mathrm{SI} + \beta_6 \cdot \mathrm{DIN} + \beta_7 \cdot \mathrm{DON}.$$

Mudeli headuse uurimisel püstitati järgmised probleemid.

Kas mudel on oluline?

Sellele küsimusele annab vastuse alljärgnev tabel:

ANOVA	df	SS	MS	F	Significance F
Regression	7	567.32	81.05	6.93	$3.95 \cdot 10^{-7}$
Residual	149	1732.05	11.70		
Total	156	2299.37			

Tabeli põhjal saame, et mudelist tingitud sinivetikate kontsentratsiooni hajuvus $S_{reg} \approx 567.32$ ning mudeli jäägist tingitud hajuvus $S_{res} \approx 1732.05$. Vastavate suuruste dimensiooniks on kontsentratsiooni ruudud. Teststatistik

$$G \approx \frac{567.3 \cdot 148}{1732 \cdot 7} \approx 6.93.$$

Teststatistiku väärtusele vastava olulisustõenäosuse (*p-value*) saame lahtrist *Significance F*. Antud suurusest *p-value* $\approx 3.95 \cdot 10^{-7}$ järeldub, et koostatud mudel on kindlasti oluline. Järelikult võime püstitada teise probleemi.

Millised mudeli parameetrid on olulised?

Mudeli kordajad $\beta_0, \beta_1, ..., \beta_7$ on leitud vähimruutude meetodil. Kuid millised neist kordajatest on olulised, millised mitte? Sellele küsimusele annab vastuse järgmine tabel:

	Coefficients	t Stat	*p-value*	Lower 95%	Upper 95%
Intercept	0.16	0.18	0.856	−1.62	1.94
BAC	−0.16	−3.40	0.000862	−0.26	−0.069
CHL	2.60	1.65	0.101	−0.52	5.71
DIP	−162.15	−3.86	0.000168	−245.14	−79.17
DOP	94.02	4.88	$2.7 \cdot 10^{-6}$	55.95	132.08
SI	−0.073	−0.15	0.880	−1.03	0.880
DIN	−2.68	−1.45	0.150	−6.34	0.98
DON	2.33	1.64	0.104	−0.48	5.15

Tabeli veeru nimega Coefficients põhjal saame kirja panna järgmise mudeli:

$$\mathrm{CY} = 0.16 - 0.16 \cdot \mathrm{BAC} + 2.6 \cdot \mathrm{CHL} - 162.15 \cdot \mathrm{DIP} + 94.02 \cdot \mathrm{DOP} - 0.073 \cdot \mathrm{SI} -$$

$$-2.68 \cdot \mathrm{DIN} + 2.33 \cdot \mathrm{DON}.$$

Järgmises veerus (t Stat) on toodud mudeli parameetritele vastavad teststatistiku väärtused. Olulised parameetrid valime olulisustõenäosuse (*p-value*) veeru põhjal. Kui olulisuse nivoo $\beta = 0.05$, siis osutuvad valituks (*p-value* $< \beta$) faktorid BAC, DIP ja DOP. Nendele faktoritele vastavate kordajate 0.95-usaldusintervalli alumine (*Lower* 95%) ja ülemine (*Upper* 95%) on sama märgiga. See tähendab, et võime olla tõenäosusega 0.95 kindlad, et vastaval faktoril on sinivetikate kontsentratsioonile kas positiivne või negatiivne mõju. Saime mudeli, mis sisaldab üksnes olulisi faktoreid

$$\mathrm{CY} = -0.16 \cdot \mathrm{BAC} - 162.15 \cdot \mathrm{DIP} + 94.02 \cdot \mathrm{DOP}.$$

Saadud tulemuse põhjal võime järeldada, et bakteroplanktonil ning mineraalsel fosforil on sinivetikate hulgale oluliselt negatiivne mõju, orgaanilise fosfori mõju aga on oluliselt positiivne.

Milline on mudeli kirjeldusvõime?

Lõpuks vastame küsimusele, kui suure osa sinivetikate kontsentratsiooni varieeruvusest kirjeldab meie koostatud mudel. Vastuse sellele annab järgmine tabel:

Regression	Statistics
Multiple R	0.497
R-Square	0.247
Adjusted R-square	0.211

Lahtrist *R-Square* saame, et determinatsioonikordaja $R^2 \approx 0.25$. Seega kirjeldab *ca* 25% sinivetikate hulga varieeruvusest mudel ning *ca* 75% mudeli jääk. Determinatsioonikordaja on küllaltki väike, mistõttu ei saa koostatud mudelisse optimistlikult suhtuda.

Kuidas aga teostada üldiste lineaarste mudelite koostamist ja diagnostikat tarkvara R abil? Demonstreerime seda äsjases näites kirjeldatud Peipsi järve andmetel. Esmalt tuleks andmetabel konverteerida .txt formaati. Siis saab andmed tarkvarasse R sisse lugeda järgmiste käskudega:

```
andmed=read.table('C:\\kataloginimi\\Peipsi.txt',
header=TRUE,
sep="\t",dec=",")
attach(andmed).
```

Andmete sisselugemiseks on R-i funktsioon

`read.table`,

mille argument
`C:\\kataloginimi\\Peipsi.txt`
näitab faili asukoha, argument `header=TRUE` käsib andmete esimese rea võtta tunnuse nimena, `sep="\t"` ütleb, et veergude eraldajaks on tabulaator ning argument `dec=","` märgib, et andmestikus on täisosa eraldajaks koma.

Lineaarne mudel ning tema väljundtabelid saadakse järgmiselt:

```
mudel=lm(CY∼BAC+CHL+DIP+DOP+SI+DIN+DON)
summary(mudel).
```

Kui aga tahame eraldada mudelist vabaliikme, siis tuleb sisestada järgmine käsk:

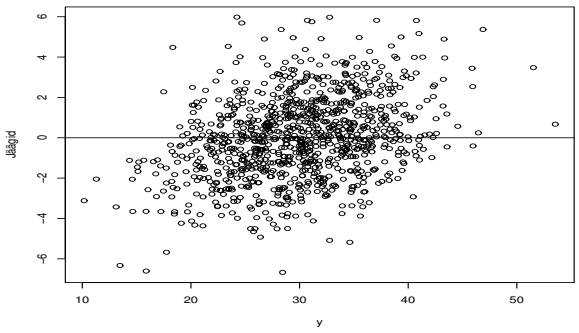
```
mudel=lm(CY∼BAC+CHL+DIP+DOP+SI+DIN+DON-1).
```

2.2.3. Mudeli jääkide analüüs

Pärast eelpoolkirjeldatud mudeli diagnostikat on tähtis analüüsida ka mudeli jääke ϵ_i, $i = 1, 2, ..., n$. Need kirjeldavad erinevust uuritava suuruse Y mudeliga saadud väärtuste ja mõõdetud väärtuste vahel. Seega

$$\epsilon_i = \widehat{y}_i - y_i.$$

Erinevuse ϵ_i näol on tegemist juhusliku suurusega. Ideaalsel juhul $\epsilon_i \sim \mathcal{N}(0, \sigma)$. See tähendab, et jääkide keskväärtus on 0 ning dispersioon ei sõltu uuritava suuruse väärtusest. Allpool oleval graafikul on toodud enam-vähem ideaalne pilt uuritava tunnuse ja mudeli jääkide vahelisest sõltuvusest.



Joonis 2.1. Uuritava suuruse Y ja mudeli jääkide ϵ_i vaheline sõltuvus

Kui graafik, mis kirjeldab jääkide ja uuritava suuruse vahelist seost, viitab mingile sõltuvusele nende vahel, siis tuleks lisada mudelisse täiendavaid faktoreid. Uuritava suuruse ja jääkide vahelise sõltuvuse ühed põhjustajad võivad olla mõjukad erindid (ingl *influential outliers*). Nendeks nimetatakse vaatlusi, mis erinevad märgatavalt teistest vaatlustest ning mõjutavad oluliselt mudelit. Uurime, kuidas hinnata seda erinevust ning erindi mõju mudelile.

Defineerime suuruse

$$\epsilon^2 = \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 = \epsilon^\intercal \epsilon.$$

Suuruse ϵ^2 näol on tegemist n vaatluse hälvete ruutude summaga. Vastaku igale vaatlusele ruuthälve $\epsilon_i^2 = (\widehat{y_i} - y_i)^2$, $i = 1, 2, ..., n$. Seda hälvet saab leida maatriksi

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$$

abil. Maatriksit $\mathbf{H} : n \times n$ nimetatakse mudeli mütsimaatriksiks (ingl *hat matrix*). Seoste (2.2) ja (2.5) põhjal saame, et

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}, \tag{2.6}$$

kus $\widehat{\mathbf{y}} = (\widehat{y_1}, \widehat{y_2}, ..., \widehat{y_n})^\intercal$ ja $\mathbf{y} = (y_1, y_2, ..., y_n)^\intercal$. Seose (2.6) põhjal saab avaldada mudeliga (2.2) saadud uuritava suuruse Y väärtused empiiriliste väärtuste kaudu järgmiselt:

$$\begin{cases} \widehat{y_1} = h_{11}y_1 + h_{12}y_2 + \ldots + h_{1n}y_n \\ \widehat{y_2} = h_{21}y_1 + h_{22}y_2 + \ldots + h_{2n}y_n \\ \vdots \\ \widehat{y_n} = h_{n1}y_1 + h_{n2}y_2 + \ldots + h_{nn}y_n. \end{cases}$$

Suuruste h_{ij} näol on tegemist mütsimaatriksi $\mathbf{H}$ elementidega. Saab näidata, et $h_{ij} = h_{ji}$ ning $0 < h_{ii} \leq 1$. Mütsimaatriksi kaudu saame avaldada hälvete ruutude summa järgmiselt:

$$\epsilon^2 = (\mathbf{y} - \mathbf{H}\mathbf{y})^\intercal(\mathbf{y} - \mathbf{H}\mathbf{y}).$$

Seega iga vaatluse ruuthälve

$$\epsilon_i^2 = \left(y_i - \sum_{k=1}^{n} h_{ik}y_k\right)^2.$$

Ruuthälvete ja maatriksi $\mathbf{H}$ abil saab mõõta i-nda vaatluse mõju lineaarsele mudelile (2.2). Küsime: kuivõrd mõjutab mudelit i-nda vaatluse kõrvaldamine? Seda mõju saab hinnata Cooki kaugusega

$$D_i = \frac{\epsilon_i^2(n-k-1)}{\epsilon^2(k+1)} \frac{h_{ii}}{(1-h_{ii})^2}. \tag{2.7}$$

Mida suurem on kauguse D_i väärtus, seda suuremat mõju avaldab mudelile i-nda vaatluse elimineerimine andmestikust.

Tarkvaras R saab tellida mudeli jäägid käsuga

`resid(mudel)`,

mille argumendiks on mudelile antud nimi. Jääkidele vastavad Cooki kaugused leiab R-i käsk

`cooks.distance(mudel)`.

Demonstreerime erindite hindamist ühe väikese näitega.

Näide 2.3. Olgu meil 5 vaatlusega andmestik

X	2	3	4	7	14
Y	3	1	5	4	19

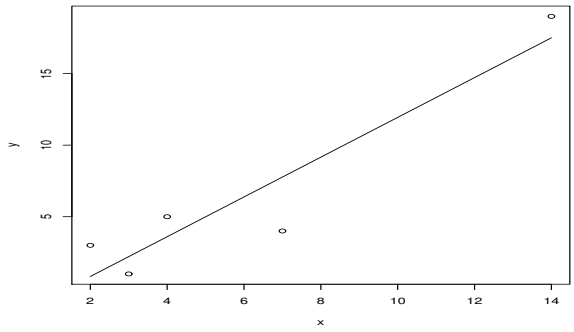
Nende andmete põhjal koostatud lineaarse mudeli olulisustõenäosus

$$p\text{-}value \approx 0.018.$$

Seega on mudel oluline, kui olulisuse nivooks võtta 0.05. Mudeli kordajaid iseloomustab järgmine tabel:

	Coefficients	p-value
Intercept	-1.96	0.44
x	1.39	0.018

Seega saime mudeli $\widehat{y_i} = -1.96 + 1.39x_i$, $i = 1, 2, ..., 5$. Visuaalselt kirjeldab seda mudelit järgnev graafik:



Joonis 2.2. Mudeliga $\widehat{y_i} = -1.96 + 1.39x_i$ saadud väärtused võrreldes empiiriliste väärtustega

Jooniselt 2.2 on näha, et üks vaatlus on ülejäänud 4 vaatlusest märgatavalt erinev. Seose (2.7) põhjal saadi järgmised Cooki kaugused D_i, mis vastasid mudeli jääkidele ϵ_i :

i	1	2	3	4	5
ϵ_i	2.17	-1.21	1.39	-3.79	1.45
D_i	0.27	0.054	0.049	0.30	7.95

On näha, et suurus D_i on teistest märgatavalt suurem, kui $i = 5$. Seega mõjutab saadud mudelit enim 5. vaatluse elimineerimine.

Testimaks vaatluse mõjukuse olulisust tuleb meil teada juhusliku suuruse D_i jaotust. Cooki kauguse jaotust on modelleeritud publikatsioonis [19]. Selles artiklis on toodud Cooki kauguse 0.95-kvantiilide tabel, mille põhjal saab leida juhusliku suuruse D_i kriitilise väärtuse olulisuse nivool 0.05.

2.3. Dispersioonanalüüsi alused

Anname ülevaate ühest väga levinud andmeanalüüsi meetodist, mida nimetatakse dispersioonanalüüsiks. Enamasti kasutatakse selle meetodi puhul ingliskeelest lühendit ANOVA (***An****alyses of* ***Va****riances*). Dispersioonanalüüs on oma olemuselt regressioonanalüüsi erijuht. Nimelt leiab ANOVA rakendust juhul, mil uuritav tunnus on pidev ning faktortunnused diskreetsed arvtunnused või koodtunnused. Nimetagem faktortunnuseid faktoriteks ning selle väärtusi faktori tasemeteks. Faktorite tasemeteks võivad näiteks olla:

1) sugu (mees või naine),

2) hoone tüübid,

3) metsa kasvukoha tüübid (puisniit, kõdusoo, kastikusoo jms).

Klassikalise dispersioonanalüüsi puhul on oluline eeldada, et uuritav tunnus allub normaaljaotusele. Käsitleme lähemalt kahte liiki dispersioonanalüüsi mudeleid.

2.3.1. Ühefaktoriline dispersioonanalüüs

Olgu uuritav tunnus Y ning faktor X. Olgu faktoril tasemed $A_1, A_2, ..., A_k$. Igal tasemel olgu m mõõtmist. Seega on meil valim

$$\mathbf{Y} = (Y_{11}, Y_{12}..., Y_{1m}, Y_{21}, ..., Y_{2m}, ..., Y_{k1}, ..., Y_{km})^\intercal,$$

mille maht $n = km$. Saame kirja panna järgmise mudeli:

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma), \quad j = 1, 2, ..., m;$$

$$\mu_i = \mu + \alpha_i, \quad i = 1, 2, ..., k.$$

Suurus μ_i tähistab mudelis uuritava suuruse keskmist i-ndal tasemel ning μ tähistab üldkeskmist. Parameetrite α_i kohta eeldatakse, et

$$\sum_{i=1}^{k} \alpha_i = 0.$$

Maatrikskujul saab dispersioonanalüüsi mudeli kirja panna järgmiselt:

$$\mathbf{y} = \mathbf{Z}\beta + \epsilon. \tag{2.8}$$

Maatriks $\mathbf{Z} : mk \times (k+1)$ seoses (2.8) kujutab endast ANOVA mudeli maatriksit, $k+1$-komponendiline vektor β koosneb suurustest

$$\mu, \alpha_1, \alpha_2, ..., \alpha_k$$

ning km-vektori ϵ elementideks on mudeli jäägid.

Näide 2.4. Olgu meil faktor X, millel on 3 taset. Igal tasemel olgu mõõtmisi 3. Siis avaldub seos (2.8) järgmiselt:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{33} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{pmatrix}.$$

Maatriksis on $\mathbf{Z}$ esimene veerg tunnuse Y keskmise veerg, ülejäänud 3 veergu aga on faktorite X erinevate tasemete veerud.

Testimaks faktori tasemete mõju püstitame dispersioonanalüüsi mudelile järgmise hüpoteeside paari:

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0, \\ H_1 : \text{leiduvad } \alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_l},\ 2 \le l \le k \text{ nii, et } \alpha_{i_1}\alpha_{i_2}...\alpha_{i_l} \neq 0. \end{cases}$$

Faktori mõju kindlakstegemine põhineb uuritava tunnuse Y hajuvuse uurimisel valimis. Jagame uuritava tunnuse hajuvuse komponentideks. Selleks leiame eraldi üldkeskmise ja i-nda faktori keskmise hinnangud

$$\overline{y} = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} \quad \text{ja} \quad \overline{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}.$$

Olgu SST (ingl ***S**um of **S**quares **T**otal*) uuritava suuruse koguhajuvus, SSB (ingl ***S**um of **S**quares **B**etween*) tasemetevaheline hajuvus ning SSW (ingl ***S**um of **S**quares **W**ithin*) tasemesisene hajuvus. Kuna

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(\overline{y}_i - \overline{y})(y_{ij} - \overline{y}_i) = \sum_{i=1}^{k}(\overline{y}_i - \overline{y})\sum_{j=1}^{m}(y_{ij} - \overline{y}_i) = 0,$$

siis

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}[(y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})]^2 =$$

$$= m\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2 = SSB + SSW.$$

Mida suurem on tasemetevaheline hajuvus SSB võrreldes tasemesisese hajuvusega SSW, seda suurem mõju on faktoril. Koostame järgnevalt teststatistiku kontrollimaks faktori mõju. On põhjust oletada, et see statistik peab sisaldama SSB ja SSW suhet.

Nullhüpoteesi korral

$$\overline{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij} \sim \mathcal{N}\left(\mu_i, \frac{\sigma}{\sqrt{m}}\right).$$

Lihtne on veenduda, et faktorite tasemete keskmine on üldkeskmine

$$\frac{1}{k}\sum_{i=1}^{k}\overline{y}_i = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} = \overline{y}.$$

Kui H_0 on õige, siis $E(\overline{y}_i) = \mu$ ning

$$\overline{y} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{km}}\right).$$

Tänu normaaljaotuse eeldusele saame lause 1.2 põhjal, et statistik

$$Q_1 = \frac{1}{\sigma^2}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 = \frac{SSB}{\sigma^2} \sim \chi^2(k-1).$$

Seega esindab Q_1 tasemetevahelist hajuvust. Edasi uurime tasemesisest hajuvust. On teada, et juhuslik suurus

$$H_i = \frac{1}{\sigma^2}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2 \sim \chi^2(m-1).$$

Kasutades χ^2-jaotuse aditiivsust (lause 1.3), saame tasemesisest hajuvust iseloomustavaks statistikuks

$$Q_0 = \sum_{i=1}^{k} H_i = \frac{SSW}{\sigma^2} \sim \chi^2(k(m-1)) = \chi^2(n-k),$$

kus $n = km$ on valimi maht. Arvestades Fisheri F-jaotuse definitsiooni, saame, et

$$G = \frac{Q_1(n-k)}{Q_0(k-1)} \sim F(k-1, n-k)$$

ehk

$$G = \frac{SSB(n-k)}{SSW(k-1)} \sim F(k-1, n-k).$$

Tuginedes seose (1.10) põhjal defineeritud G-statistiku jaotuse omadustele, loeme olulisuse nivool β sisuka hüpotesi H_1 tõestatuks ja väidame, et faktoril on mõju uuritavale tunnusele, kui

$$G > g_{1-\beta}.$$

Ülal leitud suurused on esitatud statistilise andmeanalüüsi tarkvarade väljundites enamasti järgmisel kujul:

Source	DF	Sum of Squares	Mean Square	F-Stat	$Pr > F$
Model	k-1	SSB	$SSB/(k-1)$	z	$P(Z > z)$
Error	n-k	SSW	$SSW/(n-k)$		
C Total	n-1	SST			

2.3.2. Kahefaktoriline dispersioonanalüüs

Vaatleme juhtu, kus uuritav tunnus Y sõltub faktoritest X_1 ja X_2. Olgu faktoril X_1 tasemed $A_1, A_2, ..., A_k$ ning faktoril X_2 tasemed $B_1, B_2, ..., B_h$. Antud juhul tuleb lisaks kahe faktori mõjule analüüsida ka faktorite koosmõju. Kahe faktori peale on meil kh erinevat taset. Olgu igal tasemel m mõõtmist, siis on meil valim $\mathbf{Y} = (Y_{ij1}, Y_{ij2}, ..., Y_{ijm})$, kus valimi maht $n = khm$, $i = 1, 2, ..., k$ ja $j = 1, 2, ..., h$. Valimi $\mathbf{Y}$ mingi realisatsiooni saab esitada alljärgneva tabelina, kus indeks $l = 1, 2, ..., m$:

$X_2 \setminus X_1$	A_1	A_2	$\cdots$	A_k
B_1	y_{11l}	y_{12l}	$\cdots$	y_{1kl}
B_2	y_{21l}	y_{22l}	$\cdots$	y_{2kl}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B_h	y_{h1l}	y_{h2l}	$\cdots$	y_{hkl}

Me saame kirja panna järgmise mudeli:

$$E(Y_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

Eeldame, et

$$Y_{ijl} \sim \mathcal{N}(\mu_{ij}, \sigma), \quad l = 1, 2, ...m$$

ning

$$\sum_{i=1}^{k} \alpha_i = \sum_{j=1}^{h} \beta_j = \sum_{i=1}^{k} \sum_{j=1}^{h} \gamma_{ij} = 0.$$

Antud juhul koosneb uuritava tunnuse koguhajuvus SST faktori X_1 tasemetevahelisest hajuvusest SSB_1, faktori X_2 tasemetevahelisest hajuvusest SSB_2, faktori X_1 ja X_2 tasemete kombinatsioonide vahelisest hajuvusest (ehk faktorite koosmõjust tingitud) SSB_{12} ning tasemesisesest hajuvusest SSW. Faktorite mõjude ja koosmõjude kohta saame sõnastada 3 hüpoteeside paari

1)

$$\begin{cases} H_0(X_1) : \alpha_1 = \alpha_2 = ... = \alpha_k = 0, \\ H_1(X_1) : \text{ leiduvad } \alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_l}, 2 \le l \le k \text{ nii, et } \alpha_{i_1}\alpha_{i_2}...\alpha_{i_l} \ne 0. \end{cases}$$

2)

$$\begin{cases} H_0(X_2) : \beta_1 = \beta_2 = ... = \beta_k = 0, \\ H_1(X_2) : \text{ leiduvad } \beta_{i_1}, \beta_{i_2}, ..., \beta_{i_s}, 2 \le s \le h \text{ nii, et } \beta_{i_1}\beta_{i_2}...\beta_{i_s} \ne 0. \end{cases}$$

3)

$$\begin{cases} H_0(X_1X_2) : \gamma_{11} = \gamma_{12} = ... = \gamma_{ih} = ... = \gamma_{kh} = 0, \\ H_1(X_1X_2) : \text{ leiduvad } \gamma_{i_1j_1}, \gamma_{i_2j_2}, ..., \gamma_{i_rj_r}, 2 \le r \le kh \text{ nii, et} \\ \qquad\qquad \gamma_{i_1j_1}\gamma_{i_2j_2}...\gamma_{i_rj_r} \ne 0. \end{cases}$$

Nende hüpoteeside kontrollimiseks leiame järgmiste keskmiste hinnangud:

1) tunnuse Y üldkeskmise hinnangu $\overline{y} = \dfrac{1}{khm}\sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{l=1}^{m} y_{ijk}$;

2) faktori X_1 taseme i keskmise hinnangu $\overline{y_i} = \dfrac{1}{hm}\sum_{j=1}^{h}\sum_{l=1}^{m} y_{ijk}$;

3) faktori X_2 taseme j keskmise hinnangu $\overline{y_j} = \dfrac{1}{km}\sum_{i=1}^{k}\sum_{l=1}^{m} y_{ijk}$;

4) faktori X_1 taseme i ning faktori X_2 taseme j keskmise hinnangu

$$\overline{y_{ij}} = \frac{1}{hk}\sum_{l=1}^{m} y_{ijl}.$$

Ülaltoodud hajuvusteks saame, et

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{m}(y_{ijl} - \overline{y})^2; \quad SSB_1 = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{m}(\overline{y_i} - \overline{y})^2;$$

$$SSB_2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{m}(\overline{y_j} - \overline{y})^2; \quad SSB_{12} = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{m}(\overline{y_{ij}} - \overline{y_i} - \overline{y_j} + \overline{y})^2;$$

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{m}(y_{ijl} - \overline{y_{ij}})^2 \quad .$$

Seega uuritava suuruse koguhajuvus

$$SST = SSB_1 + SSB_2 + SSB_{12} + SSW. \tag{2.9}$$

Meie eesmärk on testida tasemetevahelist hajuvust kirjeldavate liidetavate osakaalusid võrreldes liidetava SSW osakaaluga summas (2.9). Selleks koostame teststatistikud, mis alluvad Fisheri F-jaotusele. Lausete 1.2 ja 1.3 põhjal saame, et faktori X_1 mõju hindab statistik

$$G_1 = \frac{SSB_1(n-k)}{SSW(k-1)} \sim F(k-1, n-k),$$

faktori X_2 mõju hindab statistik

$$G_2 = \frac{SSB_2(n-h)}{SSW(h-1)} \sim F(h-1, n-h)$$

ning faktorite X_1 ja X_2 koosmõju tugevust testib statistik

$$G_{12} = \frac{SSB_{12}(n-kh)}{SSW(k-1)(h-1)} \sim F((k-1)(h-1), (n-kh)).$$

Antud peatükk käsitles dispersioonanalüüsi ideaalset juhtu ehk juhtu, millal uuritav tunnus Y allus normaaljaotusele ning mõõtmistulemuste hulk faktorite erinevatel tasemetel oli võrdne. Mida teha aga siis, kui need eeldused ei vasta tegelikkusele? Sellele probleemile annab vastuse järgmine peatükk.

2.3.3. Dispersioonanalüüs tarkvarade MS Excel ja R abil

Tarkvara MS Excel keskkonnas saab dispersioonanalüüsi tegemiseks kasutada protseduuri *Anova*. Selleks tuleb andmed enne salvestada tasemete kaupa ridadesse või veergudesse. Protseduur *Anova* käivitatakse töövahendist *Data Analysis*. See pakub 3 võimalust: *Single Factor*, *Two-Factor Without Replication* ning *Two-Factor With Replication*. Vaatame neid võimalusi lähemalt.

Ühefaktoriline dispersioonanalüüs

Ühefaktorilise dispersioonanalüüsi puhul valitakse käsklus *Single Factor*. Avanenud sisestusaknas tuleb määrata:

1) *Input range* – algandmete tabel;

2) *Grouped by* – määratakse, kas tasemed on orienteeritud veerge (*columns*) või ridu (*rows*) pidi;

3) *Labels in first row* – märgitakse, et esimeses reas (või veerus) on taseme nimetus;

4) *Output options* – määratakse, kuhu salvestada väljund.

Kui tasemed on orienteeritud veerge pidi, siis näeb algandmete tabel välja järgmine:

A_1	A_2	$\cdots$	A_k
y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_{1m}	y_{2m}	$\cdots$	y_{km}

Kahefaktoriline dispersioonanalüüs

Kahefaktorilise dispersioonanalüüsi puhul valitakse käsklus *Two-Factor Without Replication*. Algandmed tuleb esitada risttabelina, kus read kujutavad endast ühe ning veerud teise faktori tasemeid. Lahtri element

$$y_{ij} = \frac{1}{m} \sum_{l=1}^{m} y_{ijl},$$

$i = 1, 2, ..., k$; $j = 1, 2, ..., h$, kujutab endast aritmeetilist keskmist esimese faktori tasemel i ja teise faktori tasemel j.

Kordusmõõtmistega dispersioonanalüüs

Kordusmõõtmistega dispersioonanalüüsi puhul valitakse käsklus *Two-Factor With Replication*. Tegemist on juhuga, mil mõõtmisi on tehtud samadel objektidel eri aegadel. Sisestusaknas tuleb täiendavalt määrata

1) *Rows per sample* – kordsus (ehk mõõtmiste arv) igal ajahetkel;

2) *alpha* – olulisuse nivoo (vaikimisi 0.05).

Kordusmõõtmistega dispersioonanalüüsi puhul on algtabel järgmine:

Ajahetk 1	y_{11}	y_{21}	$\cdots$	y_{k1}
	y_{12}	y_{22}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	y_{1m}	y_{2m}	$\cdots$	y_{km}
Ajahetk 2	y_{11}	y_{21}	$\cdots$	y_{k1}
	y_{12}	y_{22}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	y_{1m}	y_{2m}	$\cdots$	y_{km}

Antud juhul on tegemist faktori tasemete hulgaga k ning kordusega m.

Dispersioonanalüüs tarkvaraga R

Mugavam on teha dispersioonanalüüsi tarkvara R abil. Selles tarkvaras saab teha dispersioonanalüüsi ka mittetasakaalustatud juhul ehk siis, kui valimi mahud erinevatel tasemetel ei ole võrdsed. Sisendtabeliks on antud juhul objekt-tunnus maatriks, kus ühes veerus on uuritava tunnuse väärtused, teises veerus faktortunnuste väärtused ehk faktorite tasemed. Pärast andmete sisselugemist saab teha tarkvara R keskkonnas dispersioonanalüüsi järgmiselt:

```
mudel=aov(Y∼X1+X2+X1*X2)
summary(mudel).
```

Dispersioonanalüüsi käsuks R-i keskkonnas on `aov`. Antud juhul on uuritavaks tunnuseks Y ning faktoriteks $X1$ ja $X2$. Nende faktorite koosmõju tähistab `X1*X2`.

2.4. Üldistatud lineaarsed mudelid

Selles osas tuleme tagasi regressioonanalüüsi probleemide juurde. Uurime juhte, mil üldised lineaarsed mudelid ei anna adekvaatseid tulemusi.

2.4.1. Eksponentsiaalsete jaotuste pere

Olgu $Y \sim F$. Antud juhul aga ei pruugi F tähendada normaaljaotust. Kuidas aga üldistada normaaljaotust? Selleks toome sisse mõiste eksponentsiaalsete jaotuste pere. Seda peret iseloomustab üldine tihedusfunktsioon

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \tag{2.10}$$

kus θ ja ϕ on parameetrid ning $a(\phi)$, $b(\theta)$ ning $c(y, \phi)$ mingid teadaolevad funktsioonid. Meie eesmärk on uurida neid parameetreid ja funktsioone erinevate jaotuste korral. Esmalt leiame eksponentsiaalsete jaotuste perele vastava logaritmilise tõepära funktsiooni $l(\theta)$. Seose (2.10) põhjal saame, et

$$l(\theta) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{2.11}$$

Osutub, et funktsiooni b kohta seoses (2.10) kehtib järgmine seaduspära.

Lause 2.3. Avaldugu juhusliku suuruse Y tihedusfunktsioon kujul (2.10). Siis

$$E(Y) = b'(\theta)$$

ja

$$D(Y) = b''(\theta)a(\phi).$$

Tõestus. Olgu meil valimi $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^\intercal$ põhjal hinnatav parameeter θ. Saab näidata, et

$$E(l'(Y, \theta)) = 0$$

ja

$$E(l''(Y, \theta)) = -D(l'(\mathbf{Y}, \theta)).$$

Leiame avaldise (2.11) liidetavas esimest ja teist järku tuletised θ järgi. Saame, et

$$\frac{\partial l(\theta,\phi)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$$

ja

$$\frac{\partial^2 l(\theta,\phi)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}.$$

Seega

$$E\left(\frac{Y - b'(\theta)}{a(\phi)}\right) = 0$$

ja

$$D\left(\frac{Y - b'(\theta)}{a(\phi)}\right) = \frac{b''(\theta)}{a(\phi)},$$

millest järeldubki, et

$$E(Y) = b'(\theta) \text{ ja } D(Y) = b''(\theta)a(\phi).$$

□

Uurime järgnevalt lähemalt, milline on konkreetsete jaotuste korral θ ja $b(\theta)$.

Normaaljaotus

Normaaljaotuse korral tihedusfunktsioon

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Esitame selle tihedusfunktsiooni kujul (2.10). Saame, et

$$f(y) = \exp\left(\ln\left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\right\}\right) =$$

$$= \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} - \ln(\sigma\sqrt{2\pi})\right) =$$

$$= \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \ln(\sigma\sqrt{2\pi})\right).$$

Antud tulemuse seosega (2.10) võrreldes saame, et $\theta = \mu$ ning $b(\theta) = \frac{1}{2}\theta^2$. Funktsiooni $a(\phi)$ väärtuseks on selle seose põhjal σ^2. Seega keskväärtus

$$E(Y) = b'(\theta) = \frac{1}{2}(\theta^2)' = \theta = \mu$$

ning dispersioon

$$D(Y) = b''(\theta)a(\phi) = \sigma^2.$$

Binoomjaotus

Olgu $Y \sim B(n, p)$. Vaatleme tihedusfunktsiooni

$$f(y) = C_n^y p^y (1-p)^{n-y}$$

kui tõenäosust $P(Y = k)$, $k =, 1, 2, ..., n$. Seda funktsiooni kujule (2.10) teisendades saame, et

$$f(y) = \exp(y\ln(p) + (n-y)\ln(1-p) + \ln(C_n^y)) =$$

$$= \exp\left(y\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln(C_n^y)\right).$$

Võrreldes saadud tulemust seosega (2.10) saame, et $a(\phi) = 1$ ning

$$\theta = \ln\left(\frac{p}{1-p}\right). \tag{2.12}$$

Seost (2.12) nimetatakse kui Logit. Avaldades sellest seosest suuruse p, saame, et

$$p = \frac{\exp(\theta)}{1+\exp(\theta)}. \tag{2.13}$$

Avaldist (2.13) nimetatakse logistiliseks seoseks. Seda seost kirjeldab graafik, mida nimetatakse sigmoidiks. Funktsioon $b(\theta)$ avaldub antud juhul kui

$$b(\theta) = n\ln(1+\exp(\theta)).$$

Seega keskväärtus

$$E(Y) = b'(\theta) = n\frac{\exp(\theta)}{1+\exp(\theta)} = np$$

ning dispersioon

$$D(Y) = b''(\theta)a(\phi) = n\frac{\exp(\theta)(1+\exp(\theta)) - \exp(2\theta)}{(1+\exp(\theta))^2} =$$

$$= n\frac{\exp(\theta)}{1+\exp(\theta)}\frac{1}{1+\exp(\theta)} = n\frac{\exp(\theta)}{1+\exp(\theta)}\left(1 - \frac{\exp(\theta)}{1+\exp(\theta)}\right) = np(1-p).$$

Poissoni jaotus

Olgu $Y \sim Po(\lambda)$. Vaatleme tihedusfunktsiooni

$$f(y) = \frac{\lambda^y}{y!}\exp -\lambda$$

kui tõenäosust $P(Y = k)$, $k = 0, 1, 2, ..., n, ...$ Teisendades selle tihedusfunktsiooni kujule (2.10), saame, et

$$f(y) = \exp(y\ln(\lambda) - \ln(y!) - \lambda).$$

Saadud tulemust avaldisega (2.10) võrreldes saame, et

$$\theta = \ln(\lambda) \text{ ehk } \lambda = \exp(\theta).$$

Funktsioon

$$b(\theta) = \exp(\theta) = \lambda.$$

Selle funktsiooni abil avalduvad juhusliku suuruse Y keskväärtus ja dispersioon järgmiselt:

$$E(Y) = b'(\theta) = \exp(\theta) = \lambda$$

ning

$$D(Y) = b''(\theta)a(\phi) = \exp(\theta) = \lambda.$$

Eksponentjaotus

Olgu $Y \sim E(\nu)$. Siis tihedusfunktsioon

$$f(y) = \begin{cases} 0, \text{ kui } y < 0, \\ \nu \exp(-\nu y), \text{ kui } y \geq 0. \end{cases}$$

Seost (2.10) rakendades saame, et $y \geq 0$ puhul

$$f(y) = \exp(\ln(\nu) - \nu y) = \exp\left(\frac{y\nu - \ln(\nu)}{-1}\right).$$

Antud juhul funktsioon $a(\phi) = -1$, parameeter

$$\theta = \nu \text{ ning } b(\theta) = \ln(\nu).$$

Seega keskväärtus

$$E(Y) = b'(\theta) = (\ln(\theta))' = \frac{1}{\theta} = \frac{1}{\nu}$$

ning dispersioon

$$D(Y) = b''(\theta)a(\phi) = -\left(\frac{1}{\theta}\right)' = \frac{1}{\nu^2}.$$

2.4.2. Seosefunktsioonid

Üldise lineaarsuse korral modelleerime seose (2.2) abil uuritava suuruse keskväärtust μ_i. Üldistame lineaarse mudeli seosega

$$\theta_i = g(\mu_i),$$

kus g on pidev ja diferentseeruv funktsioon ning

$$\theta_i = \mathbf{z}_i\beta.$$

Funktsiooni g nimetatakse seosefunktsiooniks (ingl *link function*) keskväärtuse μ ning lineaarse hinnangu θ vahel. Kuna g on pidev ja diferentseeruv, siis

$$\mu = g^{-1}(\mathbf{z}_i\beta).$$

Uurime nüüd seosefunktsioone erinevate jaotuste korral.

Tabel 2.1. Seosefunktsioonid erinevate jaotuste korral

Jaotus	Seosefunktsioon
Normaaljaotus	Ühikfunktsioon
Binoomjaotus	Logit
Poissoni jaotus	Log

Kui uuritav suurus on eksponentjaotusega, siis

$$\mu_i = E(Y_i) = \frac{1}{\lambda_i} = \frac{1}{\theta_i} = \frac{1}{\mathbf{z}_i\beta}.$$

Poissoni jaotuse puhul modelleerime mingi sündmuse sagedust. Tabelist 2.1 saame, et

$$\mathbf{z}_i\beta = \mathrm{Log}(\lambda_i)$$

ehk

$$\lambda_i = E(Y_i) = \exp(\mathbf{z}_i\beta).$$

Binoomjaotuse puhul tahame modelleerida mingi sündmuse tõenäosust p. Tabelist 2.1 saame, et

$$\mathbf{z}_i\beta = \mathrm{Logit}(p_i)$$

ehk

$$p_i = \frac{\exp(\mathbf{z}_i\beta)}{1 + \exp(\mathbf{z}_i\beta)}. \tag{2.14}$$

Mudel (2.14) kirjeldab riski või šansi sõltuvust meid huvitavatest faktoritest. Kuna antud mudel leiab küllalt laialdast rakendust, siis uurime seda lähemalt.

2.4.3. Logistilised mudelid

Uurime lähemalt olukorda, millal uuritava tunnuse Z väärtus võib olla 0 või 1. See tähendab juhtu, mil ennustame mingi sündmuse A tõenäosust. Seega on meil valimi i-nda elemendi tunnuse Z_i, $i = 1, 2, ..., n$ väärtus

$$Z_i = \begin{cases} 1, & \text{kui toimub sündmus } A, \\ 0, & \text{kui sündmust } A \text{ ei toimu.} \end{cases}$$

Juhusliku suuruse Z_i jaotust nimetatakse *Bernoulli* jaotuseks. Võtku Z_i väärtuse 1 tõenäosusega p_i ja väärtuse 0 tõenäosusega $1 - p_i$. Lihtne on veenduda, et

$$E(Z_i) = \mu_i = p_i$$

ja

$$D(Z_i) = \sigma_i^2 = p_i(1 - p_i).$$

Uurime, kuidas leida sündmuse A tõenäosuse hinnangut grupeeritud andmete korral. Jagame faktortunnuste väärtuste alusel oma n vaatlust m gruppi. Olgu igas grupis n_i vaatlust, $i = 1, 2, ..., m$. Seega

$$\sum_{i=1}^{m} n_i = n.$$

Käsitleme suurust n_i kui sõltumatute katsete hulka katsetamaks sündmuse A toimumist. Olgu juhuslik suurus

$$Y_i = \sum_{j=1}^{n_i} Z_j.$$

Seega on $Y_i = 0, 1, 2, ..., n_i$ sündmuse A toimumise suhtes õnnestunud katsete hulk igas grupis ning $Y_i \sim B(n_i, p_i)$. Järelikult

$$P(Y_i = y_i) = \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

Eelneva põhjal järeldub, et sündmuse A tõenäosuse saab leida seose (2.14) abil. Demonstreerime nüüd logistilisele mudelile vastava objekt-tunnus maatriksi koostamist ühe konkreetse andmestiku abil.

Näide 2.5. Andmestik põhineb 1975. aastal Fidžis tehtud uuringutel. Selle käigus küsitleti 1607 abielus olevat naist uurimaks rasestumisvastaste vahendite kasutamist. Küsitlust on kirjeldatud üksikasjalisemalt teadusartiklis [25]. Olgu uuritav suurus

$$Z_i = \begin{cases} 1, & \text{kui naine kasutab rasestumisvastaseid vahendeid,} \\ 0, & \text{kui naine ei kasuta rasestumisvastaseid vahendeid.} \end{cases}$$

Uuriti tõenäosuse $P(Z = 1)$ sõltuvust faktoritest X_1 – vanus, X_2 – haridus ning X_3 – soov saada rohkem lapsi. Faktorid defineeriti järgmiselt:

$$X_{1_i} = \begin{cases} a, \text{ vanus alla 25,} \\ b, \text{ vanus 25–29,} \\ c, \text{ vanus 30–39,} \\ d, \text{ vanus 40–49;} \end{cases}$$

$$X_{2_i} = \begin{cases} 0, \text{ alla keskmise haridustase,} \\ 1, \text{ üle keskmise haridustase ;} \end{cases}$$

$$X_{3_i} = \begin{cases} 0, \text{ ei soovi enam lapsi,} \\ 1, \text{ soovin veel lapsi.} \end{cases}$$

Antud 3 faktortunnuse põhjal saab 1607 naist jagada 16 gruppi. Igas grupis olgu vastavalt n_1, n_2, ..., n_{16} küsitletut. Olgu y_j, $j = 1, 2, ..., 16$, rasestumisvastaseid vahendeid tarvitavate naiste hulk grupis n_j. Gruppidevaheline jagunemine on esitatud teadusartiklis [25]. Seda jagunemist kirjeldab järgnev tabel:

X_1	X_2	X_3	y_j	n_j
a	0	1	6	59
a	0	0	4	14
a	1	1	52	264
a	1	0	10	60
b	0	1	14	74
b	0	0	10	29
b	1	1	54	209
b	1	0	27	92
c	0	1	33	145
c	0	0	80	157
c	1	1	46	164
c	1	0	78	146
d	0	1	6	41
d	0	0	48	94
d	1	1	8	16
d	1	0	31	43

Tabelist on näha, et küsitluses osales 59 alla 25 aasta vanust naist, kelle haridustase oli alla keskmise ning kes soovisid veel lapsi. Rasestumisvastaste vahendite kasutajate osakaal oli nende seas hinnanguliselt $\frac{6}{59}$. Naisi, kelle vanus oli 40–49 aastat, haridustase üle keskmise ja kes soovisid veel lapsi, oli küsitletute seas 16. Hinnanguliselt pooled neist tarvitasid rasestumisvastaseid vahendeid.

Küsime: kas faktor X_j on riskifaktor sündmuse A toimumise suhtes? Olgu meil faktor $X_j = 1$, kui tema esineb, ja $X_j = 0$, kui see faktor puudub. Vaatleme mudelit (2.14) kujul, kus X_j on ainus faktor, mis mõjutab vaatlust Z_i, $i = 1, 2, ..., n$. Siis

$$p = \frac{\exp(\beta_0 + \beta_j X_j)}{1 + \exp(\beta_0 + \beta_j X_j)}.$$

Faktori X_j olemist riskifaktoriks mõõdab šansside suhe OR (ingl ***odds ratio***). Suurus

$$\mathrm{OR} = \frac{P(Z = 1|X_j = 1)/P(Z = 0|X_j = 1)}{P(Z = 1|X_j = 0)/P(Z = 0|X_j = 0)}.$$

Logistilise mudeli põhjal saame, et

$$\mathrm{OR} = \frac{\left(\frac{\exp(\beta_0 + \beta_j)}{1 + \exp(\beta_0 + \beta_j)}\right)\left(1 - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)}{\left(1 - \frac{\exp(\beta_0 + \beta_j)}{1 + \exp(\beta_0 + \beta_j)}\right)\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)} =$$

$$= \frac{\left(\frac{\exp(\beta_0 + \beta_j)}{1 + \exp(\beta_0 + \beta_j)}\right)\left(\frac{1}{1 + \exp(\beta_0)}\right)}{\left(\frac{1}{1 + \exp(\beta_0 + \beta_j)}\right)\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)} = \frac{\exp(\beta_0 + \beta_j)}{\exp(\beta_0)} = \exp(\beta_j).$$

Seega $\beta_j > 1$ korral suurendab faktori X_j olemasolu sündmuse A toimumise tõenäosust ning $\beta_1 < 0$ korral vähendab selle faktori esinemine sündmuse A toimumise tõenäosust.

2.4.4. Üldistatud lineaarsete mudelite diagnoosimine

Järgnevalt uurime üldistatud lineaarsete mudelite parameetrite hindamist ning nende mudelite diagnostikat. Ühtlasi veendume, et tegemist on üldise lineaarse mudeli diagnostika üldistustega.

Suurima tõepära hinnang

Üldistatud lineaarsete mudelite puhul kasutatakse parameetrite vektori β hindamsel korduvalt taaskaalutud vähimruutude (ingl *iteratively re-weighted least squares*) meetodit. Selle meetodi puhul leitakse igale vaatlusele n-ö nihkemuutuja

$$z_i = \widehat{\theta}_i + (y_i - \widehat{\mu}_i)\frac{d\theta_i}{d\mu_i}$$

ning kaal

$$w_i = r_i\left[b''(\theta_i)\left(\frac{d\theta_i}{d\mu_i}\right)^2\right]^{-1},$$

kus $r_i = \dfrac{\phi}{a_i(\phi)}$. Nende näitajatega saadakse parameetrite vektorile vähimruutude hinnang

$$\widehat{\beta} = (\mathbf{Z}^\intercal\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{W}\mathbf{z},$$

kus $\mathbf{z}$ tähistab nihkemuutujate vektorit ning diagonaalmaatriksi $\mathbf{W}$: $n \times n$ elementideks on kaalud w_i.

Üldise lineaarse mudeli puhul $\theta_i = \mu_i$ ning $\dfrac{d\theta_i}{d\mu_i} = 1$. Seega $z_i = y_i$, $i = 1, 2, ..., n$. Kuna $b''(\theta_i) = 1$ ning $r_i = 1$, siis maatriks $\mathbf{W}$ on üldise lineaarse mudeli korral ühikmaatriks. Milline aga on maatriks $\mathbf{W}$ logistilise mudeli korral?

Näide 2.6. Olgu uuritav suurus μ_i mingi sündmuse tõenäosus p_i. Sellisel juhul leitakse hinnang $\widehat{p_i}$ mudeliga (2.13) ning

$$\theta_i = \ln\left(\frac{p_i}{1-p_i}\right).$$

Tuletise leidmise eeskirja alusel saame, et

$$\frac{d\theta_i}{dp_i} = \frac{1}{p_i(1-p_i)}.$$

Seega nihkemuutuja

$$z_i = \widehat{\theta}_i + \frac{y_i - p_i}{p_i(1-p_i)}.$$

Kuna $r_i = 1$ ning $b''(\theta_i) = n_i p_i(1-p_1)$, siis kaalud ehk maatriksi $\mathbf{W}$ peadiagonaali elemendid

$$w_i = \frac{p_i(1-p_i)}{n_i}.$$

Hälbimus

Võrdleme meid huvitavat mudelit küllastunud (ingl *saturated*) mudeliga. Olgu $\widehat{\theta}_i$ meid huvitava mudeli lineaarne hinnang ning $\widetilde{\theta}_i$ küllastunud mudeli lineaarne hinnang. Seosest (2.11) saame, et

$$l(\widetilde{\theta}) - l(\widehat{\theta}) = \sum_{i=1}^{n} \frac{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)}{a_i(\phi)}.$$

Tegemist on tõepära funktsioonide suhte logaritmiga, mida tähistame kui λ. Esitame tõepära suhte kui

$$2\lambda = \frac{D(\mathbf{y}, \widehat{\mu})}{\phi},$$

kus $\mathbf{y}$ tähistab mõõtmistulemuste vektorit ning

$$D(\mathbf{y}, \widehat{\mu}) = 2\sum_{i=1}^{n} r_i[y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)]. \tag{2.15}$$

Suurust $D(\mathbf{y}, \widehat{\mu})$ nimetatakse hälbimuseks (ingl *deviance*).

Näide 2.7. Vaatame juhtu, kus uuritav suurus Y on normaaljaotusega. Siis $\theta_i = \mu_i$, $b(\theta_i) = \dfrac{\theta_i^2}{2}$, $a_i(\phi) = \sigma^2$ ja $r_i = 1$. Sel juhul saame hälbimuseks

$$D(\mathbf{y}, \widehat{\mu}) = 2\sum_{i=1}^{n} \left(y_i(y_i - \widehat{\mu}_i) - \frac{y_i^2}{2} + \frac{\widehat{\mu}_i^2}{2} \right) =$$

$$= 2\sum_{i=1}^{n} \left(\frac{y_i^2}{2} - y_i\widehat{\mu}_i + \frac{\widehat{\mu}_i^2}{2} \right) = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2.$$

Seega saime, et normaaljaotuse puhul hälbimus $D(\mathbf{y}, \widehat{\mu}) = S_{res}$, mis tähendab, et hälbimus kujutab endast hälvete ruutude summat. Üldistame hälvete ruutude summa logistilisele mudelile ning Poissoni jaotusele.

Näide 2.8. Olgu $Y_i \sim B(n_i, p_i)$. Olgu küllastunud mudeli korral $\widetilde{\mu}_i = y_i$ ning olgu $\widehat{\mu}_i$ logistilise mudeli põhjal saadud STP hinnang keskväärtusele $E(Y_i) = n_i p_i$, $i = 1, 2, ..., m$. Kuna antud juhul

$$\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$$

ning

$$b(\theta_i) = n_i \ln(1 + \exp(\theta_i)),$$

siis peale mõningaid teisendusi saame, et

$$\widetilde{\theta}_i = \ln(y_i) - \ln(n_i - y_i) \text{ ja } \widehat{\theta}_i = \ln(\widehat{\mu}_i) - \ln(n_i - \widehat{\mu}_i)$$

ning

$$b(\widetilde{\theta}_i) = n_i \ln(n_i) - n_i \ln(n_i - y_i) \text{ ja } b(\widehat{\theta}_i) = n_i \ln(n_i) - n_i \ln(n_i - \widehat{\mu}_i).$$

Seose (2.15) põhjal saame hälbimuseks

$$\begin{aligned} D(\mathbf{y}, \widehat{\mu}) = 2 \sum_{i=1}^{m} \Bigg(& y_i \ln\left(\frac{y_i}{n_i}\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i}\right) - \\ & - y_i \ln\left(\frac{\widehat{\mu}_i}{n_i}\right) - (n_i - y_i) \ln\left(\frac{n_i - \widehat{\mu}_i}{n_i}\right)\Bigg) = \\ = 2 \sum_{i=1}^{m} & \left(y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right). \end{aligned}$$

Tähistades empiirilise suuruse y_i kui emp_i ning logistilise mudeli saadud suuruse $\widehat{\mu}_i$ kui teor_i, saame hälbimuse esitada kujul

$$\begin{aligned} D(\mathbf{y}, \widehat{\mu}) = 2 \sum_{i=1}^{m} \Bigg\{ & \text{emp}_i \ln\left(\frac{\text{emp}_i}{\text{teor}_i}\right) + \\ & + (n_i - \text{emp}_i) \ln\left(\frac{n_i - \text{emp}_i}{n_i - \text{teor}_i}\right)\Bigg\} \end{aligned} . \tag{2.16}$$

Näide 2.9. Olgu $Y_i \sim Po(\lambda)$. Küllastunud mudeli puhul $\widetilde{\mu}_i = y_i$. Üldiselt $\widehat{\mu}_i = \widehat{\lambda}_i$. Antud juhul

$$\theta_i = \ln(\widehat{\lambda}_i) \text{ ning } b(\theta_i) = \widehat{\lambda}_i.$$

Vastavalt Poissoni jaotuse tõenäosusfunktsioonile saame hälbimuseks

$$D(\mathbf{y}, \widehat{\lambda}) = 2\sum_{i=1}^{n}\left(y_i \ln(y_i) - y_i \ln(\widehat{\lambda}_i) - y_i + \widehat{\lambda}_i\right) =$$

$$= 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\widehat{\lambda}_i}\right) + \widehat{\lambda}_i - y_i\right).$$

Olgu $\widehat{\lambda}_i = \text{teor}_i$. Siis saame hälbimuse kujul

$$D(\mathbf{y}, \widehat{\lambda}) = 2\sum_{i=1}^{n}\left(\text{emp}_i \ln\left(\frac{\text{emp}_i}{\text{teor}_i}\right) + \text{teor}_i - \text{emp}_i\right). \tag{2.17}$$

Fisheri informatsioonikriteerium

Fisheri informatsioonikriteerium (ingl *Fisher Information Criterion*) mõõdab statistilises mudelis peituvat info hulka. Selle alusel leitakse, millist informatsiooni annab saadud parameeter θ uuritava suuruse Y kohta. Seda info hulka leitakse logaritmitud tõepärafunktsiooni $l(\theta)$ kaudu.

Definitsioon 2.2. Fisheri informatsioonikriteerium

$$I(\theta) = -E\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right). \tag{2.18}$$

Seega mõõdab statistilise mudeli informatiivsust logaritmitud tõepärafunktsiooni 2. tuletis. Järgnevalt rakendame seost (2.18) eksponentsiaalsete jaotuste perele. Seose (2.11) põhjal saame, et

$$I(\theta) = -E\left\{\frac{\partial^2}{\partial \theta^2}\sum_{i=1}^{n}\frac{Y_i\theta - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)\right\}.$$

Leiame suuruse $I(\theta)$ erinevate jaotuste puhul.

Näide 2.10. Olgu uuritav suurus $Y \sim \mathcal{N}(\mu, \sigma)$. Antud juhul on tegemist lineaarse mudeliga, kus $\theta = \mu$. Vastava logaritmitud tõepära funktsiooni põhjal saame, et

$$I(\mu) = -E\left\{ \frac{\partial^2}{\partial \mu^2} \sum_{i=1}^{n} \left(\frac{Y_i\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{Y_i^2}{2\sigma^2} - \ln(\sigma\sqrt{2\pi}) \right) \right\} =$$

$$= -E \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \mu} \left(\frac{Y_i - \mu}{\sigma^2} \right) \right\} = \frac{n}{\sigma^2}.$$

Seega on lineaarse mudeli informatiivsus võrdeline valimi mahuga ning pöördvõrdeline uuritava suuruse Y dispersiooniga.

Näide 2.11. Uurime logistilise mudeli informatiivsust. Antud juhul on uuritav suurus $Y_i \sim B(n_i, p_i)$, $i = 1, 2, ..., m$. Modelleeritavaks suuruseks on meid huvitava sündmuse tõenäosus p. Logaritmitud tõepärafunktsioon

$$l(p) = \sum_{i=1}^{m} \left(y_i \ln(p) + (n_i - y_i) \ln(1-p) + \ln(C_{n_i}^{y_i}) \right).$$

Vastavaks Fisheri informatsioonikriteeriumiks saame, et

$$I(p) = -E\left\{ \frac{\partial^2}{\partial p^2} \sum_{i=1}^{m} \left(Y_i \ln(p) + (n_i - Y_i) \ln(1-p) + \ln(C_{n_i}^{Y_i}) \right) \right\} =$$

$$= -E\left\{ \sum_{i=1}^{m} \left(\frac{\partial}{\partial p} \frac{Y_i}{p} - \frac{\partial}{\partial p} \frac{n_i - Y_i}{1-p} \right) \right\} = E\left\{ \sum_{i=1}^{m} \left(\frac{Y_i}{p^2} + \frac{n_i - Y_i}{(1-p)^2} \right) \right\} =$$

$$= \sum_{i=1}^{m} \left(\frac{n_i p}{p^2} + \frac{n_i - n_i p}{(1-p)^2} \right) = \frac{n}{p(1-p)} \sum_{i=1}^{m} n_i = \frac{n}{p(1-p)}.$$

Näide 2.12. Olgu uuritav suurus $Y \sim Po(\lambda)$. Antud juhul on modelleeritavaks suuruseks meid huvitava sündmuse keskmine sagedus λ ning logaritmitud tõepära funktsioon

$$l(\lambda) = \sum_{i=1}^{n} (y_i \ln(\lambda) - \ln(y_i!) - \lambda).$$

Selle tõepärafunktsiooni põhjal saame, et

$$I(\lambda) = -E\left\{\frac{\partial^2}{\partial\lambda^2}\sum_{i=1}^{n}(Y_i\ln(\lambda) - \ln(Y_i!) - \lambda)\right\} =$$

$$= -E\left\{\frac{\partial}{\partial\lambda}\sum_{i=1}^{n}\left(\frac{Y_i}{\lambda} - 1\right)\right\} = E\left(\sum_{i=1}^{n}\frac{Y_i}{\lambda^2}\right) = \frac{n}{\lambda}.$$

Seega, mida haruldasema sündmuse sagedust (ehk mida väiksem on λ) modelleerime, seda suurem on mudeli informatiivsus.

Fisheri informatsioonikriteeriumi kasutatakse üldistatud lineaarsete mudelite testimisel. Uurime järgnevalt neid teste.

Waldi test

Testi töötas välja Ungari statistik Abraham Wald. See võrdleb parameetri θ hinnangut $\widehat{\theta}$ meid huvitava väärtusega θ_0. Enne mudeli koostamist eeldatakse, et huvi pakkuv parameeter (mingi sündmuse sagedus, selle tõenäosus jms) on mingi kindla väärtusega. Waldi testiga uuritakse, kas koostatud mudeli põhjal saadud väärtus erineb oluliselt.

Eeldusel, et $\widehat{\theta} \sim \mathcal{N}(\theta_0, \sqrt{D(\widehat{\theta})})$ saame, et statistik

$$H = \frac{(\widehat{\theta} - \theta_0)^2}{D(\widehat{\theta})}$$

allub χ^2-jaotusele, mille vabadusastmete arv on ühemõõtmelisel juhul 1. Rakendame statistikut H üldistatud lineaarsetele mudelitele. Nende mudelite puhul $\widehat{\theta} \sim \mathcal{N}(\theta_0, \sqrt{I^{-1}(\theta)})$ ehk statistiku $\widehat{\theta}$ dispersioon võrdub Fisheri informatsioonikriteeriumiga. Seega statistik

$$H = (\widehat{\theta} - \theta_0)^2 I(\theta).$$

Seda statistikut kasutatakse testimaks hüpoteeside paari

$$\begin{cases} H_0 : \theta = \theta_0, \\ H_1 : \theta \neq \theta_0. \end{cases}$$

Näide 2.13. Rakendame Waldi testi logistilisele mudelile. Vaatame hüpoteeside paari mingi sündmuse tõenäosusele p

$$\begin{cases} H_0 : p = 0.5, \\ H_1 : p \neq 0.5. \end{cases}$$

Logistilise mudeliga saadi 20 vaatluse põhjal, et $\widehat{p} = 0.62$. Kuna Fisheri informatsioonikriteerium $I(p) = \dfrac{n}{p(1-p)}$, siis saadi H_0 kehtimise eeldusel teststatistiku väärtuseks

$$h = (0.62 - 0.5)^2 \frac{20}{0.5(1-0.5)} = 1.152.$$

Vabadusastmete arvu 1 puhul saame antud väärtusele h vastava olulisustõenäosuse *p-value* ≈ 0.28. Seega oleme kohustatud jääma nullhüpoteesi H_0 juurde.

Skoori test

Teine võimalus testimaks, et $\widehat{\theta} = \theta_0$, on kasutada skoori testi. Selleks defineeritakse suurus

$$u(\theta) = \frac{\partial \ln(L(\theta))}{\partial \theta}.$$

Mudeli leitud STP hinnangu $\widehat{\theta}$ puhul $u(\widehat{\theta}) = 0$. Testimaks hüpoteeside paari

$$\begin{cases} H_0 : \theta = \theta_0, \\ H_1 : \theta \neq \theta_0 \end{cases}$$

koostatakse teststatistik

$$S = \frac{u(\theta)^2}{I(\theta)}.$$

Ühemõõtmelisel juhul on statistik S asümptootiliselt χ^2-jaotusega vabadusastmete arvuga 1. Logistilise mudeli puhul

$$u(p) = \frac{1}{p}\sum_{i=1}^{m} y_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{m} y_i\right) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)},$$

kus k on mudeli põhjal saadud sündmuse sagedus. Statistik

$$S = \frac{(k - np)^2}{np(1-p)}.$$

Akaiki informatsioonikriteerium

Järgnevalt uurime statistiliste mudelite teist informatsioonikriteeriumit, mida iseloomustab suurus AIC (ingl ***A****kaike* ***I****nformation* ***C****riterion*). Eesti keeles on selle suuruse pikem nimetus Akaiki informatsioonikriteerium. See mõõdab koostatud mudeli kvaliteeti. Kriteerium ise avaldub kui

$$\mathrm{AIC} = 2k - \ln(L(\theta)),$$

kus $L(\theta)$ tähistab mudeli tõepära funktsiooni maksimaalset väärtust ning k mudelis olevate parameetrite hulka. Suurus AIC pärineb informatsiooniteooriast. Tegemist on entroopia ehk korrapäratuse mõõduga. See suurus on väga oluline kriteerium valimaks paljude mudelite seast parimat. Selle valiku põhimõte on: *kõikvõimalike mudelite seast tuleb valida selline mudel, millele vastav suurus AIC on vähim.* Suuruse AIC kõrge väärtus annab põhjust kahtlustada, et mudel on üleparametriseeritud. Liiga suur hulk parameetreid annab mänguruumi manipuleerimisteks.

2.4.5. Üldistatud lineaarsete mudelite koostamine ning diagnoosimine tarkvara R abil

Tarkvaras R on üldistatud lineaarsete mudelite analüüsiks käsk `glm`. Mudeli koostamiseks ja diagnoosimiseks tuleb koostada järgmine programm:

```
mudel=glm(y∼x1+x2+x3,family=)
summary(mudel),
```

milles `y` tähistab uuritavat tunnust ning `x1`, `x2` ja `x3` on faktortunnuste tähistused. Argumendiga `family` määratakse mudeli tüüp ning seosefunktsioon. Mõningaid näiteid:

1) üldise lineaarse funktsiooni korral
`family=gaussian(link = "identity");`

2) logistiliste mudelite korral
`family=binomial(link = "logit");`

3) kui on eesmärk modelleerida uuritava suuruse keskväärtuse pöördväärtust, siis
`family=Gamma(link = "inverse");`

4) meid huvitava sündmuse sageduse modelleerimise korral
`family=poisson(link = "log")`.

Logistilise mudeli puhul on uuritav suurus `y` kas väärtustega 0 või 1. Teine võimalus (seda just grupeeritud andmete korral) on koostada logistiline mudel käsuga

`mudel=glm(cbind(y,n−y)` ∼`x1+x2+x3,family=binomial)`,

kus `n` tähistab sõltumatute katsete hulka ja `y` mingi sündmuse toimumise suhtes õnnestunud katsete hulka.

Toome mõningad näited üldistatud lineaarsete mudelite uurimisest.

Näide 2.14. Vaatleme andmestikku, millega uuritakse ühe füüsilise testi sooritamise tõenäosuse sõltuvust erinevatest teguriteks. Nendeks teguriteks on pikkus, kinganumber ning testi sooritaja kuulumine kas esimesse või teise rühma. Uuritava tunnuse väärtuseks on 1, kui test sooritati, ning 0, kui testi ei sooritatud. Koostati 7 mudelit ning igale mudelile leiti vastav suurus AIC. Tulemused olid järgmised:

Tegurid	AIC
Pikkus	119.86
Kinganumber	120.6
Rühm	115.01
Pikkus, Kinganumber	120.92
Kinganumber, Rühm	116.95
Pikkus, Rühm	116.98
Pikkus, Kinganumber, Rühm	118.38

Tabelist on näha, et suuruse AIC vähim väärtus vastas mudelile, kus oli ainsaks teguriks (ehk faktortunnuseks) tunnus Rühm. Ühtlasi osutus see tegur kõikide mudelite puhul ka ainsaks, millele vastav $p - value < 0.05$. Faktortunnust Rühm sisaldav mudel osutus järgmiseks:

$$y = \frac{\exp(-1.49 + 1.09x)}{1 + \exp(-1.49 + 1.09x)},$$

kus

$$x = \begin{cases} 1, \text{ kui testi sooritaja kuulus 1. rühma,} \\ 0, \text{ kui testi sooritaja kuulus 2. rühma.} \end{cases}$$

Seega suurendas 1. rühma kuulumine testi sooritamise tõenäosust, kusjuures šansside suhe $OR = \exp(1.09)$.

Näide 2.15. Järgnev näide põhineb teadusartiklil [2]. Selles uuriti riski, et hoone hallituse indeks (ingl *mould index*) tõuseb üle lubatavuse piiri. Olgu selleks ebasoodsaks sündmuseks A. Selle tõenäosuse kirjeldamiseks saadi logistiline mudel

$$P(A) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{6} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{6} \beta_i X_i\right)}, \tag{2.19}$$

mis koosnes järgmistest faktortunnustest:

1) X_1 – isolatsioonikihi paksus (mm);

2) X_2 – soojusjuhtivus (W·K/m);

3) X_3 – sisetemperatuur (°C);

4) X_4 – niiskuslisa (g/m^3);

5) X_5 – ülejäänud kihi paksus (mm);

6) X_6 – aurutõke (0/1 tunnus).

Võrdlemaks erinevate dimensioonidega suurusi viidi läbi järgmine standardiseerimine:

$$X_i = \begin{cases} -1, \text{ kui } X_i < \mu_i - \sigma_i, \\ 0, \text{ kui } X_i \in [\mu_i - \sigma_i; \mu_i + \sigma_i], \\ 1, \text{ kui } X_i > \mu_i + \sigma_i, \end{cases}$$

$i = 1, 2, ..., 5$. Suurus μ_i tähistab i-nda faktortunnuse keskväärtust ning σ_i selle tunnuse dispersiooni. Pannes standardiseeritud suurused X_1-X_5 mudelisse (2.19) saadi tarkvara R abil parameetritele β_j, $j = 0, 1, ..., 6$ järgmised hinnangud:

Parameeter	Hinnang	*p-value*
β_0	-2.578	$1.87 \cdot 10^{-8}$
β_1	6.373	$3.56 \cdot 10^{-13}$
β_2	-6.704	$2.78 \cdot 10^{-13}$
β_3	-2.429	$1.59 \cdot 10^{-8}$
β_4	4.051	$1.8 \cdot 10^{-11}$
β_5	-4.757	10^{-9}
β_6	-2.01	$2.75 \cdot 10^{-7}$

Tabelis olevatest olulisustõenäosustest (*p-value*) järeldub, et mudeli kordajad on kõik statistiliselt olulised. Samuti on tabelist näha (seda vastava kordaja märgist), et tunnustel X_2, X_3, X_5 ja X_6 on sündmuse A tõenäosusele pärssiv mõju ning tunnustel X_1 ja X_4 soodustav mõju.

2.5. Faktoranalüüs

See osa käsitleb teatud valdkondades levinud andmeanalüüsi meetodit, mida nimetatakse faktoranalüüsiks. Meetodit rakendatakse valdkondades, kus mõõtmistulemused koosnevad väga paljudest tunnustest ning eesmärgiks on need tunnused süstematiseerida. Nendeks valdkondadeks on näiteks

1) keskkonnauuringud, kus andmestik sisaldab palju erinevaid näitajaid;

2) meetriline antropoloogia, kus tuleb erinevad tunnused jagada klassidesse;

3) testid ja sotsioloogilised uuringud, mille põhjal tahetakse analüüsida tunnuseid, mida otseselt ei saa mõõta.

Faktoranalüüsi näol on tegemist andmeanalüüsi meetodiga, mis eeldab maatriksalgebraalaseid teadmisi. Uurime järgnevalt faktoranalüüsi kahest vaatenurgast.

2.5.1. Tunnuste jagamine faktoriteks

Olgu meil k tunnusega mõõtmistulemuste vektor $\mathbf{X} = (X_1, X_2, ..., X_k)^\intercal$. Vektori $\mathbf{X}$ komponentide seas, mis on ülejäänutega tugevasti seotud (korrelatsioonikordajate absoluutväärtused on suured). Võib öelda, et osa tunnuseid kirjeldavad praktiliselt üht ja sama asja ehk n-ö dubleerivad üksteist. Praktiliselt võib kogu informatsiooni juhusliku vektori $\mathbf{X}$ kohta saada tunnustega $Y_1, Y_2, ..., Y_r$, $r < k$. Nimetagem neid tunnuseid faktoriteks. Antud juhul mõistame faktorite all tunnuseid, mida tahame uurida, kuid neid pole võimalik otseselt mõõta. Näiteks on psühholoogidel eesmärk uurida loogilist mõtlemist, loovust ja verbaalset võimekust. Neid aga ei saa otseselt mõõta, vaid tuleb leida erinevate testide abil.

Iga komponent X_i on ligikaudselt esitatav faktorite lineaarkombinatsioonina järgmiselt:

$$X_i \approx \sum_{j=1}^{r} a_{ij} Y_j, \quad i = 1, 2, ..., k.$$

Maatrikskujul avaldub see lineaarne kombinatsioon kui

$$\mathbf{X} \approx \mathbf{AY}. \tag{2.20}$$

Maatriksit $\mathbf{A} : k \times r$ seoses (2.20) nimetatakse faktorlaadungite maatriksiks, maatriksit $\mathbf{Y} : r \times 1$ aga faktorite vektoriks. Lihtsustamaks faktoritesse jagamise ülesannet võib esitada faktoritele Y_j mõned lisatingimused. Seda ilma ülesande püstitust kitsendamata. Need tingimused on järgmised.

1° Eeldame, et kõikide faktorite keskväärtused on nullid:

$$E(Y_j) = 0, \quad j = 1, 2, ..., r.$$

2° Faktorid on paarikaupa mittekorreleeruvad, s.t

$$E(Y_i Y_j) = E(Y_i)E(Y_j) = 0, \quad i \neq j, \quad i, j = 1, 2, ..., r.$$

3° Faktorite dispersioonid võrduvad ühega:

$$D(Y_j) = 1, \quad j = 1, 2, ..., r.$$

Seose (2.20) täpne kuju avaldub kui

$$\mathbf{X} = \mathbf{AY} + \mathbf{U}, \tag{2.21}$$

kus vektorit $\mathbf{U} = (U_1, U_2, ..., U_k)^\intercal$ nimetatakse omapärade (ingl *uniqueness*) vektoriks. Seega on iga tunnus X_i esitatav faktorlaadungite lineaarkombinatsiooni ja omapära U_i summana. Mida aga kujutab endast faktorlaadungi maatriksi element a_{ij}? Sellele annab vastuse järgmine lause.

Lause 2.4. Faktorlaadungite maatriksi element a_{ij} on tunnuse X_i ja faktori Y_j vaheline lineaarne korrelatsioonikordaja, s.t

$$a_{ij} = \mathrm{corr}(X_i, Y_j).$$

Tõestus. Üldisust kitsendamata võime tunnused $X_1, X_2, ..., X_k$ standardiseerida. Andes faktoritele eespool toodud lisatingimused, saame, et

$$\mathrm{corr}(X_i, Y_j) = \frac{\mathrm{cov}(X_i, Y_j)}{\sqrt{D(X_i)}\sqrt{D(Y_j)}} = E(X_i Y_j) = \sum_{l=1}^{r} E\{a_{il} Y_l + U_i) Y_j\} =$$

$$= \sum_{l=1}^{r} a_{il} E(Y_l Y_j) + E(U_i Y_j) = a_{ij} D(Y_j) = a_{ij}.$$

□

Tekib küsimus, kui suure osa tunnuse X_i hajuvusest kirjeldab mudel (2.20) ning kui suure osa omapära U_i. Faktoranalüüsi mudelile allumist kirjeldab suurus

$$h_i^2 = \sum_{j=1}^{r} a_{ij}^2.$$

Suurust h_i^2 nimetatakse tunnuse X_i kommunaliteediks. Pärast tunnuse X_i standardiseerimist

$$1 = D(X_i) = \sum_{j=1}^{r} a_{ij}^2 + D(U_i).$$

Seega

$$0 \le h_i^2 \le 1 \text{ ning } 0 \le D(U_i) \le 1, \quad i = 1, 2, ..., k.$$

Järelikult kirjeldab tunnuse X_i hajuvusest $h_i^2 \cdot 100\%$ faktoranalüüsi mudel (2.20) ning $(1 - h_i^2) \cdot 100\%$ omapära U_i. Toome sisse suuruse

$$l_j = \sum_{i=1}^{k} a_{ij}^2.$$

Suurus l_j iseloomustab faktori Y_j efektiivsust. See efektiivsus näitab, kui suure osa vektori $\mathbf{X}$ koguvarieeruvusest kirjeldab faktor Y_j. Kuna

$$\sum_{i=1}^{k} E(X_i^2) = \sum_{i=1}^{k} D(X_i) = k = \sum_{i=1}^{k} \sum_{j=1}^{r} a_{ij}^2 + \sum_{i=1}^{k} E(U_i^2),$$

siis kehtivad alati võrratused

$$0 \le \sum_{j=1}^{r} l_j \le k.$$

Suurus $l = \sum_{j=1}^{r} l_j$ iseloomustab, kui suures osas kirjeldab faktorite komplekt $Y_1, Y_2, ..., Y_r$ vektori $\mathbf{X}$ hajuvust. Selleks osaks on $\frac{l}{k} \cdot 100\%$.

Faktoriteks jagamine tarkvara R abil

Vaatame, kuidas koostatada mudelit (2.21) tarkvara R abil.

Näide 2.16. Tuleme tagasi Peipsi järve andmestiku juurde näites 2.2. Jagame selles andmestikus 8 mõõtmistulemust 2 faktoriks. Pärast andmestiku sisselugemist saab seda teha järgmiste R-i käskudega:

```
x=cbind(CY,BAC,CHL,DIP,DOP,SI,DIN,DON)
factanal(x,factors=2)
```

Seega faktoritesse jagab käsk `factanal`, milles tuleb märkida andmestik $\mathbf{X}$ ning soovitud faktorite hulk $r = 2$. Saadi järgmised väljundid:

`Uniquenesses`

CY	BAC	CHL	DIP	DOP	SI	DIN	DON
0.91	0.887	0.93	0.251	0.005	0.005	0.854	0.735

`Loadings`

	Factor 1	Factor 2
CY	0.287	0
BAC	0.322	0
CHL	0.224	0.140
DIP	0.856	0.126
DOP	0.997	0
SI	0	0.995
DIN	0	0.378
DON	0.514	0

`Kirjeldusvõime`

	Factor 1	Factor 2
Proportion Var	0.279	0.148

Tabelis `Uniquenesses` olevad suurused näitavad, kui suure osa tunnuse X_i hajuvusest kirjeldab omapära U_i ehk ligikaudsele mudelile (2.20) allumatuse määra. Antud juhul alluvad faktoriteks jagamisele kõige rohkem orgaaniline fosfor (DOP) ning räni (SI). Kõige omapärasemateks (ehk faktoriteks jagamisele mittealluvateks) tunnusteks aga on sinivetikas (CY) ja klorofüll (CHL).

Tabelis `Loadings` on toodud faktorlaadungite maatriksi $\mathbf{A}$ struktuur. Selle tabeli alusel toimub tunnuste paigutamine faktorite alla. Tunnus sobib enim selle faktori alla, mille faktorlaadungi absoluutväärtus on suurim. Tabelist järeldub, et faktori 1 (Y_1) alla sobivad enim tunnused DIP ja DOP. Faktori 2 (Y_2) alla käib aga tunnus räni (SI). Seega võiks faktorit 1 nimetada fosfori faktoriks, faktorit 2 aga räni faktoriks.

Kolmas tabel näitab, kuivõrd kirjeldavad faktorid 8 tunnuse varieeruvust. Järeldub, et faktor 1 kirjeldab tunnuste hajuvust 27.9% ning faktor 2 osa selles on 14.8%. Kokku on seega 2 faktori kirjeldusvõime 42.7%.

Faktoriteks jagamist saab teha ka tunnuste $X_1, X_2, ..., X_k$ korrelatsioonimaatriksi $\mathbf{R}$ baasil. Demonstreerime seda järgnevalt kümnevõistluse

andmestikul.

Näide 2.17. Uurime faktoranalüüsi andmestikul, mis koosneb kümnevõistluse erinevate suurvõistluste tulemustest. Tunnuste nimed andmestikus on järgmised:

v1 – 100 m jooks, v2 – kaugushüpe, v3 – kuulitõuge, v4 – kõrgushüpe, v5 – 400 m jooks, v6 – 110 m tõkkejooks, v7 – kettaheide, v8 – teivashüpe, v9 – odavise, ja v10 – 1500 m jooks.

Algselt oli teada nende tunnuste vaheline korrelatsioonimaatriks. Jagamaks selle korrelatsioonimaatriksi põhjal 10 ala faktoriteks koostati järgmine R-i programm:

```
x=cbind(v1,v2,v3,v4,v5,v6,v7,v8,v9,v10)
ability.cov=x
factanal(factors=4,covmat=ability.cov).
```

Väljundiks saadi faktorlaadungite tabel

Ala	Factor 1	Factor 2	Factor 3	Factor 4
v1	0.167	0.857	0.246	−0.178
v2	0.239	0.476	0.581	
v3	0.963	0.153	0.201	
v4	0.242	0.172	0.632	0.113
v5		0.71	0.236	0.331
v6	0.205	0.261	0.588	
v7	0.699	0.133	0.179	
v8	0.138		0.512	0.177
v9	0.418		0.175	
v10			0.113	0.988

Selle tabeli põhjal võib kümnevõistluse alad jagada faktoriteks järgnevalt:

Spordiala	Faktori nr
100 m jooks	2
Kaugushüpe	2–3
Kuulitõuge	1
Kõrgushüpe	3
400 m jooks	2
110 m tõkkejooks	3
Kettaheide	1
Teivashüpe	3
Odavise	1
1500 m jooks	4

Sellest tulenevalt võks faktoreid nimetada järgmiselt: faktor 1 – jõud, faktor 2 – kiirus, faktor 3 – hüppevõime või osavus ning faktor 4 – vastupidavus.

Faktoriteks jagamine on küllaltki tinglik ning teatud määral ka loominguline. See eeldab paljuski koostööd vastavate alade (sotsioloogia, antropoloogia, keskkonnatehnoloogia jms) spetsialistidega.

2.5.2. Peakomponentide meetod

Uurime nüüd faktoranalüüsi vaatenurgast, mida nimetatakse peakomponentide meetodiks (ingl *principle component method*). Tuleme tagasi seoste (2.20) ning (2.21) juurde. Esitame nendest mudelitest lähtudes seose tunnuste $X_1, X_2, ..., X_k$ vahelise korrelatsioonimaatriksi $\mathbf{R}$ ning faktorlaadungite maatriksi $\mathbf{A}$ vahel:

$$\mathbf{R} = E(\mathbf{X}\mathbf{X}^\intercal) = E(\mathbf{A}\mathbf{Y}\mathbf{Y}^\intercal\mathbf{A}^\intercal) = \mathbf{A}E(\mathbf{Y}\mathbf{Y}^\intercal)\mathbf{A}^\intercal.$$

Faktoritele esitatud lisatingimuste põhjal saame, et maatriks $E(\mathbf{Y}\mathbf{Y}^\intercal)$ on faktorite $Y_1, Y_2, ..., Y_r$ kovariatsioonimaatriks, mis on diagonaalmaatriks. Tähistagem seda maatriksit kui $\mathbf{\Lambda}$. Seega saame võrduse

$$\mathbf{R} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\intercal. \tag{2.22}$$

Peakomponentide meetod seisneb võrrandi (2.22) lahendamises maatriksi $\mathbf{A}$ ning diagonaalmaatriksi $\mathbf{\Lambda}$ suhtes. Samuti on eesmärgiks leida faktorite

Y_j avaldised tunnuste X_i kaudu ehk leida selline maatriks $\mathbf{B} : r \times k$, et

$$\mathbf{Y} = \mathbf{BX}.$$

Avaldame maatriksi $\mathbf{B}$ korrelatsioonimaatriksi $\mathbf{R}$ kaudu:

$$\Lambda = \mathbf{BRB}^\intercal. \tag{2.23}$$

Peakomponentide ehk faktorite leidmine taandub võrrandi (2.23) lahendamisele maatriksite $\mathbf{\Lambda}$ ning $\mathbf{B}$ suhtes. Kuna maatriks $\mathbf{B}$ ei ole võrrandiga (2.23) üheselt määratud, siis lisatakse ülesandele lisatingimused:

1) maatriksi $\mathbf{B}$ read olgu normeeritud, s.t

$$\sum_{i=1}^{k} b_{ji}^2 = 1, \quad j = 1, 2, ..., r;$$

2) peakomponendid Y_j olgu järjestatud dispersioonide kahanemise suunas, s.t

$$\lambda_j \geq \lambda_{j+1} \quad j = 1, 2, ..., r-1.$$

Esimene peakomponent tuleb valida nii, et tema dispersioon oleks maksimaalne. Selleks leiame maatriksi $\mathbf{B}$ reavektori $\mathbf{b}_1^\intercal = (b_{11}, b_{12}, ..., b_{1m})$ selliselt, et

$$\mathbf{b}_1^\intercal \mathbf{R} \mathbf{b}_1 = \max_{\mathbf{b}} \mathbf{b}^\intercal \mathbf{R} \mathbf{b}.$$

Saime ekstreemumülesande vektori $\mathbf{b}_1$ suhtes, kuhu tuleb lisada ortonormeerituse tingimus

$$\mathbf{b}_1^\intercal \mathbf{b}_1 = 1.$$

Kokkuvõttes jõudsime tingliku ekstreemumi leidmise ülesandeni, mida lahendadakse Lagrange'i määramata kordajate meetodil ehk n-ö λ-meetodil. Selle ülesande lahendamiseks moodustatakse funktsioon

$$\phi(\mathbf{b}) = \mathbf{b}^\intercal \mathbf{R} \mathbf{b} - \lambda(\mathbf{b}_1^\intercal \mathbf{b}_1 - 1),$$

mis tuleb maksimeerida vektori $\mathbf{b}$ suhtes. Selleks võtame funktsioonist $\phi(\mathbf{b})$ tuletise vektori $\mathbf{b}$ järgi ning võrdsustame selle nulliga:

$$\frac{\partial \phi(\mathbf{b})}{\partial \mathbf{b}} = 2\mathbf{Rb} - 2\lambda \mathbf{b} = 0.$$

Jõudsime tunnuste $X_1, X_2, ..., X_k$ vahelise korrelatsioonimaatriksi $\mathbf{R}$ omaväärtuste ja omavektorite probleemini

$$(\mathbf{R} - \lambda \mathbf{I}_k)\mathbf{b} = \mathbf{0}, \tag{2.24}$$

kus $\mathbf{I}_k$ tähistab $k \times k$ ühikmaatriksit ning $\mathbf{0}$ nullvektorit. Lausest 2.1 järeldub, et korrelatsioonimaatriksi $\mathbf{R} : k \times k$ omaväärtused on mittenegatiivsed ning nende summa $\sum_{i=1}^{k} \lambda_i = k$. Liidetavate osas esineb 2 äärmust:

1) kui kõik tunnused $X_1, X_2, ..., X_k$ on omavahel mittekorreleeruvad, siis $\lambda_1 = \lambda_2 = ... = \lambda_k = 1$;

2) kui kõigi tunnuste $X_1, X_2, ..., X_k$ vaheliste korrelatsioonikordajate absoluutväärtus on 1, siis $\lambda_1 = k$ ning $\lambda_i = 0$, $i = 2, 3, ..., k$.

Esimesel juhul on andmestiku iga tunnus üks komponent. Teisel juhul aga koosneb andmestik ühest komponendist. Meid huvitab vahepealne variant. Vaatame olukorda, kus maatriksi $\mathbf{R}$ nullist erinevate omaväärtuste hulk $r < k$.

Esmalt leiame korrelatsioonimaatriksi $\mathbf{R}$ suurima omaväärtuse λ_1 ning paneme selle võrrandis (2.24) suuruse λ asemele. Sellega saame täpse eeskirja leidmaks suurimale omaväärtusele vastavat omavektorit $\mathbf{b}_1$. Pärast selle leidmist saame, et

$$Y_1 = \mathbf{b}_1^{\intercal}\mathbf{X} \text{ ning } D(Y_1) = \lambda_1.$$

Komponent Y_1 kirjeldab suurima osa tunnuste vektori $\mathbf{X}$ varieeruvusest. Järgnevalt leiame suuruselt järgmisele omaväärtusele λ_2 vastava omavektori $\mathbf{b}_2$. Seda tingimusel, et see vektor oleks ortogonaalne vektoriga $\mathbf{b}_1$:

$$\mathbf{b}_2^{\intercal}\mathbf{b}_1 = 0.$$

Omavektori $\mathbf{b}_2$ abil saame teise peakomponendi Y_2. Analoogiliselt jätkates saame leida maatriksi $\mathbf{R}$ omaväärtuste ja omavektorite abil kõik peakomponendid.

Maatriksi $\mathbf{B}$ abil saab leida faktorlaadungite maatriksi $\mathbf{A}$. See käib järgnevalt. Arvestades, et maatriksi $\mathbf{B}$ reavektorid on ortonormeeritud ning üksteisega ortogonaalsed, saame, et

$$\mathbf{B}\mathbf{B}^{\intercal} = \mathbf{I}_r \text{ ning } \mathbf{B}^{\intercal} = \mathbf{B}^{-1}.$$

Sellest järeldub, et

$$\mathbf{B}^\intercal\mathbf{Y} = \mathbf{B}^\intercal\mathbf{B}\mathbf{X} = \mathbf{I}_r\mathbf{X} = \mathbf{X}.$$

Seosest (2.20) omakorda järeldub, et

$$\mathbf{A} = \mathbf{B}^\intercal.$$

Seega määrab maatriks $\mathbf{B}$ ka faktorlaadungite maatriksi.

Vaatame, kuidas realiseerida peakomponentide meetodit tarkvara R abil.

Näide 2.18. Rakendame peakomponentide Peipsi järve andmestiku 8 tunnusele tarkvara R käsuga

```
summary(princomp(x,cor=TRUE))
```

järgmise väljundtabeli:

C1	C2	C3	C4	C5	C6	C7	C8
1.69	1.26	1.00	0.86	0.8	0.78	0.65	0.42
0.356	0.198	0.124	0.091	0.08	0.075	0.054	0.022
0.356	0.554	0.678	0.769	0.849	0.924	0.978	1

Tabelis on kirjeldatud andmestiku komponente C1–C8. Tabeli esimeses reas on komponentide standardhälbed $\sqrt{D(Y_j)}$, $j = 1, 2, ..., 8$. Teises ja kolmandas reas toodud arvud näitavad, kuivõrd kirjeldab iga komponent andmestikus olevat 8 tunnust. Neist järeldub, et komponendid C1, C2 ja C3 kirjeldavad koguvarieeruvusest vastavalt 35.6%, 19.8% ja 12.4%. Kõigi komponentide koguvarieeruvus on 67.8%.

Viimastel aastatel on peakomponentide meetod leidnud laialdast rakendust biomeditsiinis ja geneetikas. Sellega on näiteks modelleeritud inimese mingi geeni alleelide sageduse varieeruvust erinevates maailma osades.

Peakomponentide meetodi rakendamisel geneetikas kasutatakse sageli geenimarkereid (ingl *genetic marker*). Geenimarkerite näol on tegemist geeni või DNA (desoksüribonukleiinhappe) mingi lõiguga, mille asukoht on kromosoomis teada. See lõik on lihtsalt määratava DNA struktuuriga, mis

on genoomis teadlast huvitavale geenile piisavalt lähedal, et koos selle geeniga meioosis järglastele päranduda. Enim kasutatud geenimarkerid on ühenukleotiidilised DNA polümorfismid (ingl *single nucleotide polymorphism*). Seega kujutab geenimarker endast küllaltki mugavat viisi, kuidas mõõta geenialleelide sagedusi populatsioonis. Samuti saab geenimarkeritega identifitseerida liiki või indiviidi.

Teadusartiklis [32] on kasutatud bialleelseid geenimarkereid rakendamaks peakomponentide meetodit populatsiooni struktuuri uurimisel. Selles publikatsioonis olev andmestik on esitatud maatriksina $\mathbf{C}$, mille rida tähistab indiviidi ning veerg markerit. Kokku on uuritud n indiviidi ja m markerit. Järelikult on maatriksi $\mathbf{C}$ näol tegemist $n \times m$-maatriksiga, mille element c_{ij} tähistab alleeli. Igale veerule ehk markerile on vastavusse seatud keskmise vektor

$$\mu_j = \frac{\sum_{i=1}^{n} c_{ij}}{n}, \quad j = 1, 2, ..., m.$$

Siis saadakse n-ö tsentreeritud marker

$$\overline{c}_{ij} = c_{ij} - \mu_j.$$

Võttes $p_j = \frac{\mu_j}{2}$, saadakse iga indiviidi igale alleelile suurus

$$m_{ij} = \frac{\overline{c}_{ij}}{\sqrt{p_j(1-p_j)}}.$$

Suurus m_{ij} iseloomustab i-nda indiviidi j-markeri normeeritud sagedust. Nendest sagedustest saadakse maatriks $\mathbf{M}$. Järgnevalt leitakse $n \times n$-maatriks

$$\mathbf{X} = \frac{1}{n}\mathbf{M}\mathbf{M}^{\intercal},$$

millel põhineb populatsiooni geneetilise struktuuri uurimine. Maatriksi $\mathbf{X}$ omaväärtuste ja omavektorite alusel jagatakse populatsioon komponentideks.

Kokkuvõttes võib öelda, et peakomponentide meetodi rakendamine geneetikas on muutumas üha laialdasemaks.

2.6. Ülesanded

Ülesanne 2.1. Olgu tunnustel X_1 ja X_2 järgmised väärtused:

X_1	5	7	8	11	4	3	2
X_2	6.2	7.8	9.8	9.4	6.4	6.8	6.3

Leidke lineaarne korrelatsioonikordaja $\mathrm{corr}(X_1, X_2)$.

Ülesanne 2.2. Olgu meil järgmised andmed:

X	1	4.1	3.2	5	7	2.2
Y	4	17.2	15.3	17.1	19.3	5.3

Nende andmete põhjal koostati empiiriline valem $Y = \beta_0 + \beta_1 X$. Leidke parameetritele β_0 ja β_1 hinnangud vähimruutude meetodil.

Ülesanne 2.3. Olgu meil alljärgnevas tabelis uuritava suuruse Y empiirilised väärtused ning mudeliga $Y = 5 + 2X$ saadud teoreetilised väärtused.

Empiiriline	20.8	22.1	33.4	19.8	13.8	23.0	26.6	21.8	21.6
Teoreetiline	22.1	20.8	28.7	22.5	8.6	27.2	25.6	25.4	22.0

Leidke determinatsioonikordaja R-ruut. Mis järelduse teete sellest kordajast?

Ülesanne 2.4. Eeldati, et valguse neeldumist iseloomustav ekstinktsioonikoefitsient Y ($\mathrm{M}^{-1}\mathrm{cm}^{-1}$) sõltub mingi aine molaarsest kontsentratsioonist (M) järgmiselt:

$$Y = \beta_0 + \beta_1 X.$$

Mõõtmisel saadi järgmine tulemuste tabel:

X	0.4	0.7	1.0	1.2	1.4	1.7	2.0
Y	0.23	0.34	0.42	0.55	0.61	0.77	0.86

1) Kui palju muutub keskmiselt ekstinktsioonikoefitsient, kui aine kontsentratsioon tõuseb 1 ühiku võrra?

2) Milline on mudeli põhjal saadud ekstinktsioonikoefitsient aine 0.5 molaarse kontsentratsiooni korral?

3) Leidke ülaltoodud lineaarse mudeli determinatsioonikordaja R^2.

Ülesanne 2.5. Olgu meil uuritav tunnus Y ning faktortunnuste vektor $\mathbf{X} = (X_1, X_2)^\intercal$. Nende tunnuste vahelisi seoseid iseloomustavad järgmised korrelatsioonikordajad:

$$\mathrm{corr}(Y, X_1) = 0.15, \quad \mathrm{corr}(Y, X_2) = 0.6 \text{ ning } \mathrm{corr}(X_1, X_2) = -0.2.$$

Leidke mitmene korrelatsioonikordaja $\mathbf{r}$.

Ülesanne 2.6. On antud järgmine andmestik, milles on 12 objekti:

Y(g)	X_1(m)	X_2(cm)	X_3(m^2)	X_4(cm)
51.4	0.2	17.8	24.6	18.9
72	1.9	29.4	20.7	8
53.2	0.2	17	18.5	22.6
83.2	10.7	30.2	10.6	7.1
57.2	6.8	15.3	8.9	27.3
66.5	10.6	17.6	11.1	20.8
98.3	9.6	35.6	10.6	5.6
74.8	6.3	28.2	8.8	13.1
92.2	10.8	34.7	11.9	5.9
97.1	9.6	35.8	10.8	5.5
88.1	10.5	29.6	11.7	7.8
94.8	20.3	26.6	6.7	10.1

1) Koostage üldine lineaarne mudel $Y = \beta_0 + \sum_{j=1}^{4} \beta_j X_j$.

2) Testige koostatud mudeli olulisust olulisuse nivool 0.05.

3) Milliste parameetrite β_j puhul saab kummutada hüpoteesi $H_0 : \beta_j = 0$, $j = 0, 1, ..., 4$? Olulisuse nivooks võtta 0.05.

4) Milline on mudeli kirjeldusvõime?

Ülesanne 2.7. Erinevatel kuudel mõõdeti täiskasvanud hirvede toidu tarbimist (kg) päevas. Saadi järgmised tulemused:

Veebruar	Mai	August	November
4.7	4.6	4.8	4.9
4.9	4.4	4.7	5.2
5	4.3	4.6	5.4
4.8	4.5	4.4	5.1
4.7	4.1	4.7	5.6

Kas võib nende tulemuste põhjal väita, et täiskasvanud hirvede toidu tarbimine sõltub aastaajast? Olulisuse nivooks võtta 0.05.

Ülesanne 2.8. Järvevee hapnikusisaldust (mg/l) mõõdeti 3 erineva meetodiga. Saadi järgmised tulemused:

Meetod 1	Meetod 2	Meetod 3
10.96	10.88	10.73
10.77	10.81	10.78
10.9	10.8	10.79
10.69	10.81	10.82
10.87	10.7	10.88
10.6	10.82	10.81

Kas võib olulisuse nivool 0.05 ümber lükata väite, et meetod ei mõjuta mõõtmistulemust?

Ülesanne 2.9. Taheti kontrollida noorukite *fitness*-testi tulemuse sõltuvust proteiinirikka toidu söömisest ja soost. Selleks teostati *fitness*-test 10 neiu ja 10 noormehe peal. Kusjuures pooltele anti eelnevalt proteiinirikast, pooltele aga proteiinivaest toitu. Saadi järgmised tulemused 10-punktisüsteemis:

Sugu	Kõrge proteiin	Madal proteiin
Mees	10	5
	7	7
	9	4
	5	4
	8	5
Naine	5	3
	4	4
	6	5
	3	1
	2	2

Kas testi tulemus sõltus rohkem soost või proteiini tarbimisest? Kas soo ja proteiini tarbimise koosmõju on oluline? Olulisuse nivooks olgu 0.05.

Ülesanne 2.10. Leidke Poissoni jaotusele vastav nihkemuutuja z_i ning maatriks $\mathbf{W}$.

Ülesanne 2.11. Koostati mudel hindamaks Poissoni jaotuse parameetrit λ. Selle parameetri hinnanguks saadi 30 vaatluse põhjal 1.3. Kas võib olulisuse nivool 0.05 kummutada hüpoteesi, et $\lambda = 1$? Kontrollimiseks kasutada Waldi või skoori testi.

Ülesanne 2.12. Koostati mudel, et hinnata eksponentjaotuse parameetrit ν. Selle parameetri hindamiseks koostati 25 vaatluse põhjal statistiline mudel, mis andis uuritava tunnuse aritmeetiliseks keskmiseks $\overline{x} = 1.5$. Kas võib olulisuse nivool 0.05 kummutada hüpoteesi, et $\nu = 1$? Kontrollimiseks kasutada Waldi või skoori testi.

Ülesanne 2.13. Olgu meil uuritav suurus Y_i, $i = 1, 2, ..., n$ eksponentjaotusega parameetriga ν. Koostati mudel hindamaks keskväärtust $E(Y_i)$. Leidke selle mudeli hälbimus kujul (2.16) või (2.17).

Ülesanne 2.14. Allugu uuritav suurus Y geomeetrilisele jaotusele, $Y \sim Geo(p)$. Seega

$$P(Y = y) = p(1 - p)^{y-1},$$

$y = 1, 2, ..., n, ...$ Veenduge, et tegemist on eksponentsiaalsete jaotuste peresse kuuluva jaotusega. Milline on vastav seosefunktsioon?

Ülesanne 2.15. Allugu uuritav suurus Y diskreetsele Pareto jaotusele (ülesanne 1.6). Leidke lahendused järgmistele probleemidele.

1) Veenduge, et diskreetne Pareto jaotus kuulub eksponentsiaalsete jaotuste peresse.

2) Leidke seose (2.10) kordaja θ ning funktsioon $b(\theta)$ diskreetse Pareto jaotuse puhul.

3) Leidke funktsiooni b abil keskväärtus $E(Y)$.

4) Koostage mudel modelleerimaks parameetrit a.

Ülesanne 2.16. Uuriti, mitu tundi kulutati kontrolltööks valmistumiseks. Uurimus viidi läbi 12 tudengi peal. Pärast vaadati, kas nad said kontrolltöö arvestatud (arvestus = 1) või mitte (arvestus = 0). Saadi järgmised tulemused:

Tunnid	0.5	0.75	1	1.25	1.5	1.75	2	2.5	3	3.5	4	5
Arvestus	0	0	0	0	0	1	1	1	0	0	1	1

1) Koostage mudel, mis kirjeldab tõenäosust $p = P(\text{arvestus} = 1)$.

2) Kas arvestuse soorituse tõenäosus sõltus oluliselt kontrolltööks valmistumiseks kulunud ajast, kui olulise nivooks võtta 0.05?

3) Kontrollige olulisuse nivool 0.05 hüpoteesi $p = 0.5$.

Ülesanne 2.17. Koostage näite 2.5 baasil logistilised mudelid, mis kirjeldavad abielunaiste rasestumisvastaste vahendite kasutamise tõenäosuse sõltuvust naise vanusest, haridustasemest ja tema soovist saada veel lapsi. Võrrelge erinevate mudelite puhul suurust AIC.

Ülesanne 2.18. Koosnegu 45 silda kirjeldav andmestik järgmistest tunnustest: AEG – ehituseks kulunud aeg päevades; PINDALA – silla pindala (m^2); MAKSUMUS – silla ehituse kulu eurodes; ARVUKUS – silda puudutavate jooniste hulk; PIKKUS – silla pikkus meetrites; SPAN – sillete arv; KEERUKUS – silla konstruktsiooni keerukus (0 – lihtne ja 1 – keeruline).

Kõik need 7 tunnust jagati 2 faktoriks. Saadi järgmise struktuuriga faktorlaadungite maatriks:

Tunnus	Faktor 1	Faktor 2
AEG	0.687	0.465
PINDALA	0.759	0.445
MAKSUMUS	0.831	0.351
ARVUKUS	0.592	0.648
PIKKUS	0.927	0.122
SPAN	0.864	0.201
KEERUKUS	0.165	0.937

1) Paigutage 7 tunnust erinevate faktorite alla. Kuidas võiks nimetada faktorit 1 ja faktorit 2?

2) Leidke igale tunnusele vastavad kommunaliteedid.

3) Kui suure osa 7 tunnuse hajuvusest kirjeldavad need 2 faktorit?

Ülesanne 2.19. Olgu meil tunnuste vektor $\mathbf{X} = (X_1, X_2, X_3)^\intercal$. Miks ei saa olla nende tunnuste vaheliseks korrelatsioonimaatriksiks maatriks

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}?$$

Ülesanne 2.20. Millised allpool toodud maatriksitest võivad olla mingite tunnuste vaheliseks korrelatsioonimaatriksiks?

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & -0.8 \\ -1 & 1 & -1 \\ -0.8 & -1 & 1 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 1 & -0.9 & -0.1 \\ -0.9 & 1 & -0.1 \\ -0.1 & -0.1 & 1 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 1 & -0.2 & 0.95 & -0.5 \\ -0.2 & 1 & 0.1 & 0.5 \\ 0.95 & 0.1 & 1 & -0.2 \\ -0.5 & 0.5 & -0.2 & 1 \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}.$$

Näpunäide: kasutage peamiinorite meetodit.

Ülesanne 2.21. Tunnuste X_1, X_2, X_3, X_4, X_5 ning X_6 vaheline korrelatsioonimaatriks

$$\mathbf{R} = \begin{pmatrix} 1 & 0.47 & 0.73 & -0.47 & -0.04 & 0.66 \\ 0.47 & 1 & 0.87 & -0.24 & -0.49 & 0.85 \\ 0.73 & 0.87 & 1 & -0.54 & -0.39 & 0.84 \\ -0.47 & -0.24 & -0.54 & 1 & -0.32 & -0.29 \\ -0.04 & -0.49 & -0.39 & -0.32 & 1 & -0.23 \\ 0.66 & 0.85 & 0.84 & -0.29 & -0.23 & 1 \end{pmatrix}.$$

Jagada maatriksi $\mathbf{R}$ alusel tunnused komponentideks. Kasutada peakomponentide meetodit. Mida sellest järeldada?

3. peatükk
Mitteparameetriline statistika

Eelmises kahes peatükis järgiti statistiliste mudelite koostamisel järgmisi aspekte:

1) uuritava tunnuse jaotus on teada;

2) modelleeriti uuritava suuruse keskväärtust.

Selles peatükis läheneme statistilisele modelleerimisele teisest vaatenurgast. Vaatenurgast, mida nimetatakse mitteparameetriliseks statistikaks. Uurime juhtumeid, kus üldkogumi jaotuse kohta info puudub või on see puudulik. Samuti vaatame, kuidas modelleerida juhusliku suuruse maksimaalseid ja minimaalseid väärtusi. Esmalt aga toome välja peamised erinevused klassikalise ja mitteparameetrilise statistika vahel.

3.1. Mitteparameetrilise statistika erinevused võrreldes klassikalise statistikaga

Klassikaline statistika põhineb suuresti parameetrilisuse eeldusel, mis seisneb teadmises, et valimivektor $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ on üldkogumist, mis allub teadaolevale jaotusele $F(\Theta)$. Meie eesmärgiks oli hinnata selle jaotuse parameetrite vektorit $\Theta = (\theta_1, \theta_2, ..., \theta_k)^\intercal$. Antud peatükis loobume parameetrilisuse eeldusest ehk eeldusest üldkogumi jaotusest.

Peamised erinevused klassikalise ja mitteparameetrilise statistika vahel on toodud järgmises tabelis.

Tabel 3.1. Klassikalise ja mitteparameetrilise statistika erinevusi

Klassikaline statistika	**Mitteparameetriline statistika**
Tunnus peab olema järjestatav	Saab rakendada ka koodtunnusele
Eeldab uuritava tunnuse jaotust	Puudub eeldus uuritava tunnuse jaotusest
Eeldab suurt valimi mahtu	Valimi maht võib olla väike
Kasutab andmestiku toorinfot	Kasutab andmestiku modifitseeritud infot

Tabeli 3.1 põhjal võib järeldada, et klassikaline statistika on jaotustekeskne, mitteparameetriline statistika aga on jaotuste eeldusest vaba. Klassikalises statistikas koostasime teadaoleva jaotuse parameetri θ_j hindamiseks statistiku $\widehat{\theta_j} = T(\mathbf{X})$. Selles peatükis loobume jaotuse parameetri juures indeksist. Meil pole ju eeldusi jaotusest, seega puuduvad eeldused jaotuse parameetrite hulgast. Mitteparameetrilisel juhul modifitseerime esmalt oma valimit teisendusega $\mathbf{X}^* = g(\mathbf{X})$. Seejärel leiame statistiku $T(\mathbf{X}^*)$. Üks levinumaid andmestiku modifikatsioone on teisendus, kus juhusliku suuruse väärtuse asemel kasutatakse tema astakut (ingl *rank*).

3.2. Ekstremaalsete väärtuste teooria

Seni oleme uurinud keskväärtuse modelleerimist (keskmine suurus, sündmuse keskmine tõenäosus, keskmine esinemissagedus jms). Sageli aga ei huvita meid keskmised väärtused, vaid ekstremaalsed väärtused. Juhuslike suuruste ekstremaalsete väärtuste teooria on tähtis haru matemaatilises statistikas. Seda kasutatakse tugevusõpetuses, meditsiinis, keskkonnateadustes jne. Teooria aitab prognoosida näiteks järgneva 30 aasta jõgede maksimaalset ja minimaalset vooluhulka või mingi ravimi mõju maksimaalse vererõhu languse hindamisel. Spordis on rakendatud ekstremaalsete väärtuste teooriat ennustamaks tuleviku rekordeid. Ka kindlustuses pakuvad huvi pigem ekstremaalsed kui keskmised väärtused.

3.2.1. Järkstatistikute jaotused

Ekstremaalsete väärtuste teooria põhineb järkstatistikute ning nende jaotuste uurimisel. Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$. Defineerime järkstatistikud kui

$$X_{(1)} = \min_{1 \le i \le n} \{X_i\}, \quad X_{(2)} = \min_{\substack{1 \le i \le n \\ X_{(1)} \ne X_{(2)}}} \{X_i\}, \ldots, \quad X_{(n)} = \max_{1 \le i \le n} \{X_i\}.$$

Seega

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Seda loetelu nimetatakse teoreetiliseks variatsiooni reaks. Leiame selle rea maksimaalse ja minimaalse liikme jaotused. Maksimaalse liikme jaotusfunktsioon

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(X_1 \le x, X_2 \le x, ..., X_n \le x) =$$

$$= \prod_{i=1}^{n} P(X_i \le x) = F^n(x).$$

Vastav tihedusfunktsioon

$$f_{X_{(n)}}(x) = nF^{n-1}(x)f(x).$$

Minimaalse liikme puhul saame jaotusfunksiooniks

$$F_{X_{(1)}}(x) = P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) =$$

$$= 1 - P(X_1 > x, X_2 > x, ..., X_n > x) = 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - (1 - F(x))^n$$

ning vastavaks tihedusfunktsiooniks

$$f_{X_{(1)}}(x) = n(1 - F(x))^{n-1} f(x).$$

Milline aga on i-nda järkstatistiku $X_{(i)}$ jaotusfunktsioon, $i = 1, 2, ..., n$? Rakendades binoomjaotust, saame järkstatistikute $X_{(k)}$ ning $X_{(k+1)}$ puhul, et

$$P(X_{(k)} < x \le X_{(k+1)}) = C_n^k F^k(x)(1 - F(x))^{n-k}.$$

Kõikide nende tõenäosuste summeerides osutub variatsioonirea üldliikme ehk juhusliku suuruse $X_{(i)}$ jaotusfunktsiooniks

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} \frac{n!}{k!(n-k)!} F^k(x)(1-F(x))^{n-k}, \quad i = 1, 2, ..., n.$$

Vastava tihedusfunktsiooni $f_{X_{(i)}}(x)$ avaldise saame jaotusfunktsiooni diferentseerimisel:

$$f_{X_{(i)}}(x) = \sum_{k=i}^{n} \frac{n!}{k!(n-k)!} (kF^{k-1}(x)(1-F(x))^{n-k} -$$

$$-(n-k)F^k(x)(1-F(x))^{n-k-1}) f(x).$$

Paneme tähele, et summa viimases liikmes võrdub lahutatav nulliga ja samuti seda, et

$$\frac{n!}{(k+1)!(n-(k+1))!}(k+1)F^{k+1-1}(1-F(x))^{n-(k+1)} -$$

$$-\frac{n!}{k!(n-k)!}(n-k)F^k(x)(1-F(x))^{n-k-1} = 0.$$

Seega koonduvad liidetavad vastastikku ning alles jääb üksnes summa esimese liidetava plussiga liige. Järelikult saame i-nda järkstatistiku tihedusfunktsiooniks

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}(1-F(x))^{n-i} f(x).$$

Järkstatistiku $X_{(i)}$ keskväärtus

$$E(X_{(i)}) = \int_{-\infty}^{\infty} x f_{X_{(i)}}(x)dx =$$

$$= \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{\infty} xF(x)^{i-1}(1-F(x))^{n-i} f(x)dx.$$

Püstitame küsimuse, kuidas leida tõenäosust, et juhusliku suuruse minimaalne ja maksimaalne väärtus kuuluvad lõiku $[x_1; x_2]$. See tähendab, et peame leidma $P(X_{(1)} > x_1, X_{(n)} \leq x_2)$. Arvestades, et valimisse kaasatud juhuslikud suurused on sõltumatud, saame, et

$$P(X_{(1)} > x_1, X_{(n)} \leq x_2) =$$

$$= P(x_1 < X_1 \le x_2, x_1 < X_2 \le x_2, ..., x_1 < X_i \le x_2, ..., x_1 < X_n \le x_2) =$$

$$\prod_{i=1}^{n} P(x_1 < X_i \le x_2) = \{F(x_2) - F(x_1)\}^n.$$

Saadud tulemusest järeldub, et valimi mahu suurendamine vähendab tõenäosust, et minimaalne ja maksimaalne väärtus kuuluvad mingisse kindlasse lõiku.

Näide 3.1. Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$, kus X_i allub normaaljaotusele $\mathcal{N}(\mu, \sigma)$. Siis i-nda järkstatistiku tihedusfunktsioon

$$f_{X_{(i)}}(x) = \frac{n!}{(2\pi\sigma^2)^{\frac{n}{2}}(i-1)!(n-i)!}\left\{\int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt\right\}^{i-1} \times$$

$$\times\left\{1 - \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt\right\}^{n-i} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Laplace'i veafunktsiooni kaudu avaldub i-nda järkstatistiku tihedusfunktsioon järgmiselt:

$$f_{X_{(i)}}(x) = \frac{n!}{\sqrt{2\pi}(i-1)!(n-i)!}\left\{0.5 + \Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{i-1} \times$$

$$\times\left\{0.5 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{n-i} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Näide 3.2. Olgu meil valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$, kus X_i allub eksponentjaotusele parameetriga $\nu = 1$. Siis saame minimaalse väärtuse $X_{(1)}$ jaotusfunktsiooniks

$$F_{X_{(1)}} = P(X_{(1)} \le x) = 1 - \exp(-nx)$$

ning tihedusfunktsiooniks

$$f_{X_{(1)}}(x) = n\exp(-nx).$$

Maksimaalse väärtuse $X_{(n)}$ jaotus- ja tihedusfunktsioon on aga järgmise kujuga:

$$F_{X_{(n)}} = P(X_{(n)} \le x) = (1 - \exp(-x))^n$$

ning

$$f_{X_{(n)}}(x) = n(1-\exp(-x))^{n-1}\exp(-x).$$

Järkstatistiku $X_{(i)}$, $i = 1, 2, ..., n$ jaotusfunktsioon

$$P(X_{(i)} \le x) = \sum_{k=i}^{n} \frac{n!}{i!(n-i)!}(1-\exp(-x))^i \exp(-(n-i)x)$$

ning tihedusfunktsioon

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}(1-\exp(-x))^{i-1}\exp(-(n-i+1)x).$$

3.2.2. Ekstremaalsete väärtuste jaotused

Uurime, mis saab järkstatistikute jaotustega, kui valimi maht $n \to \infty$. Teisisõnu, meie eesmärk on leida juhuslike suuruste ekstremaalsetele väärtustele vastavad jaotused. Vaatame seda ühel küllaltki laial erijuhul. Olgu meil juhuslikud suurused $X_1, X_2, ..., X_n$ sõltumatud ja sama jaotusega (s.s.j). Olgu

$$Y_n = \max\{X_1, X_2, ..., X_n\}, \quad n \ge 1.$$

Kuna

$$\max\{X_1, X_2, ..., X_n\} = -\min\{-X_1, -X_2, ..., -X_n\},$$

siis võib väita, et maksimaalsete väärtuste teooria on identne minimaalsete väärtuste teooriaga ning vastupidi.

Definitsioon 3.1. Olgu $X_1, X_2, ..., X_n, ...$ s.s.j juhuslikud suurused. Olgu $Y_n = \max_{1\le k\le n}\{X_k\}$, $n \ge 1$. Öeldakse, et juhusliku suuruse X jaotusfunktsioon kuulub samasse klassi jaotusfunktsiooniga G, kui eksisteerivad sellised jadad $\{a_n > 0, n \ge 1\}$ ja $\{b_n, n \ge 1\}$, mille korral juhuslik suurus

$$X = \frac{Y_n - b_n}{a_n}$$

koondub jaotuse järgi jaotusfunktsiooniks G.

Uurime lähemalt, milline on see jaotuste klass. Selleks sõnastame ühe teoreemi.

Teoreem 3.1. Definitsiooni 3.1 tingimusi rahuldavad järgmised 3 jaotusfunktsiooni:

Fréchet' jaotus

$$G(x) = \begin{cases} 0, \text{ kui } x < 0, \\ \exp(-x^{-\alpha}), \text{ kui } x \geq 0, \end{cases} \quad \alpha > 0;$$

Weibulli jaotus

$$G(x) = \begin{cases} \exp(-(-x)^{\alpha}), \text{ kui } x < 0, \\ 1, \text{ kui } x \geq 0, \end{cases} \quad \alpha > 0;$$

ning Gumbeli jaotus

$$G(x) = \exp\Big(- e^{-x} \Big).$$

Teoreemi 3.1 väites esitatud jaotusi nimetatakse ekstremaalsete väärtuste jaotusteks. Selle teoreemi põhjaliku tõestuse võib huviline leida teadusartiklist [14]. Demonstreerime paari näitega ekstremaalsete väärtuste jaotuste tekkimist.

Näide 3.3. Olgu juhuslikud suurused $X_1, X_2, ..., X_n, ...$ s.s.j, kus $X_n \sim \mathcal{E}(1)$. Olgu meil $Y_n = \max\{X_1, X_2, ..., X_n\}$, $n \geq 1$. Siis jaotusfunktsioon

$$F(x) = \begin{cases} 0, \text{ kui } x < 0, \\ 1 - e^{-x}, \text{ kui } x \geq 0 \end{cases}$$

ning

$$P(Y_n \leq x) = (1 - e^{-x})^n.$$

Tuntud piirväärtuse

$$\lim_{n\to\infty} \left(1 - \frac{u}{n}\right)^n = e^{-u}$$

põhjal valida definitsioonis 3.1 toodud jadad järgmiselt: $a_n = 1$ ja $b_n = \ln n$. Saame, et

$$\lim_{n\to\infty} P(Y_n \leq x + \ln n) = \lim_{n\to\infty} (1 - e^{-x-\ln n})^n = \lim_{n\to\infty} \left(1 - \frac{e^{-x}}{e^{\ln n}}\right)^n =$$

$$= \lim_{n\to\infty}\left(1-\frac{e^{-x}}{n}\right)^n = \exp\left(-e^{-x}\right).$$

Seega tekkis Gumbeli jaotus.

Näide 3.4. Olgu meil juhuslikud suurused $X_1, X_2, ..., X_n, ...$ s.s.j, kus X_n allub ühtlasele jaotusele lõigus $[0; a]$. Olgu meil $Y_n = \max\{X_1, X_2, ..., X_n\}$, $n \geq 1$. Siis jaotusfunktsioon

$$F(x) = \begin{cases} 0, \text{ kui } x < 0, \\ \dfrac{x}{a}, \text{ kui } x \in [0; a], \\ 1, \text{ kui } x > a \end{cases}$$

ning

$$P(Y_n \leq x) = \left(\frac{x}{a}\right)^n.$$

On ilmne, et juhuslik suurus Y_n koondub tõenäosuse järgi suuruseks a. Seega on meil sobilikum uurida juhuslikku suurust $a - Y_n$ Saame, et

$$P(a - Y_n \leq x) = P(Y_n \geq a - x) = 1 - \left(\frac{a-x}{a}\right)^n = 1 - \left(1 - \frac{x}{a}\right)^n.$$

Antud juhul on sobilik valida definitsioonis 3.1 toodud jadad järgmiselt: $a_n = \dfrac{1}{n}$ ja $b_n = a$. Siis saame $x < 0$ korral, et

$$\lim_{n\to\infty} P(n(Y_n - a) \leq x) = \lim_{n\to\infty} P\left(a - Y_n \geq \frac{-x}{n}\right) = \lim_{n\to\infty}\left(1 - \frac{-x}{an}\right)^n =$$

$$= \exp\left(-\frac{-x}{a}\right).$$

Seega tekkis Fréchet' jaotus.

3.3. Tunnustevahelised astakkorrelatsioonid

Klassikalise statistika peatükis käsitletud Pearsoni korrelatsioonikordajal on mitmeid puuduseid. Nendest peamised on järgmised:

1) korrelatsioonikordaja iseloomustab üksnes juhuslike suuruste vahelist lineaarset sõltuvust;

2) on tundlik erindite suhtes;

3) annab ebaadekvaatset infot grupeeritud andmete korral.

Uurime Pearsoni korrelatsioonikordaja käitumist erinevate sõltuvuste korral. Tekitame selleks juhuslike arvude generaatoriga 1000-elemendilise valimi, mille iga element X_i, $i = 1, 2, ..., 1000$, on koopia juhuslikust suurusest $X \sim \mathcal{N}(0, 1)$. Alljärgnevalt on toodud seosed juhuslike suuruste X ja Y vahel ning sellele seosele vastav korrelatsioonikordaja $\text{corr}(X, Y)$:

Seos	Pearsoni korrelatsioonikordaja
$Y = \exp(-X)$	$\text{corr}(X, Y) \approx -0.77$
$Y = X^2$	$\text{corr}(X, Y) \approx 0$
$Y = X^3$	$\text{corr}(X, Y) \approx 0.78$

Kõigil kolmel juhul on olemas kindel seos (ja sõltuvus) juhuslike suuruste X ja Y vahel, kuid Pearsoni korrelatsioonikordaja ei kirjelda seda täielikult. Kui lineaarne korrelatsioonikordaja ei anna adekvaatset tulemust, siis tuleb kasutada korrelatsioonikordajaid, mis põhinevad tunnuse astakutel (ingl *rank*).

Definitsioon 3.2. Juhusliku suuruse X väärtuse x_i astakuks nimetatakse tema järjekorranumbrit variatsioonireas $x_1, x_2, ..., x_i, ..., x_n$.

Vähim järjekorranumber 1 vastab vähimale väärtusele, suurim järjekorranumber n suurimale väärtusele. Olgu meil näiteks variatsioonirida

$$(1, 2.1, 0.4, 3, 7).$$

Siis vastavad astakud on $(2, 3, 1, 4, 5)$.

3.3.1. Spearmanni kordaja

Uurime järgnevalt korrelatsioonikordajat, mida nimetakse Spearmanni astakkorrelatsioonikordajaks. Olgu meil vaatlused (x_1, y_1), (x_2, y_2), ... ja (x_n, y_n) esitatud alljärgneva tabelina:

Y	Y astak	X	X astak
y_1	q_1	x_1	r_1
y_2	q_2	x_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	q_i	x_i	r_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	q_n	x_n	r_n

Selles tabelis tähistavad $r_1, r_2, ..., r_n$ tunnuste $x_1, x_2, ..., x_n$ astakuid ja $q_1, q_2, ..., q_n$ tunnuste $y_1, y_2, ..., y_n$ astakuid. Defineerime suuruse

$$D = \sum_{i=1}^{n} (q_i - r_i)^2.$$

Uurime astakute paare (r_i, q_i), $i = 1, 2, ..., n$, kahel äärmuslikul juhul.

Esimese äärmuse puhul on juhuslikud suurused X ja Y seotud monotoonselt kasvavalt. Sel juhul $q_i = r_i$ iga $i = 1, 2, ..., n$ puhul. Siis saame järgmised astakute paarid:

$$(1, 1), (2, 2), ..., (n, n).$$

Sel juhul suurus $D = 0$.

Teise äärmuse korral on juhuslikud suurused X ja Y seotud monotoonselt kahanevalt. Siis on astakute paarid järgmised:

$$(1, n), (2, n-1), ..., (n, 1).$$

Sel juhul suurus

$$D = \sum_{i=1}^{n} \Big(i - (n - i + 1)\Big)^2 = \sum_{i=1}^{n} \Big(2i - (n+1)\Big)^2 =$$

$$= \sum_{i=1}^{n} \Big(4i^2 - 4i(n+1) + (n+1)^2\Big) =$$

$$= \frac{2}{3} n(n+1)(2n+1) - 2n(n+1)^2 + n(n+1)^2 =$$

$$= n\left(\frac{2}{3}(n+1)(2n+1) - (n+1)^2\right) =$$

$$= n\left(\frac{1}{3}n^2 - \frac{1}{3}\right) = \frac{n(n^2-1)}{3} := S.$$

On ilmne, et suurus $D \geq 0$. Järgnevalt leiame suurusele D ülemise piiri.

Lause 3.1. Kehtib võrratus

$$D \leq S.$$

Tõestus. Lihtne on veenduda, et monotoonselt kahaneva seose puhul astak

$$r_i = n + 1 - q_i. \tag{3.1}$$

Olgu meil suvalised astakud

$$(1, \widetilde{q_1}), (2, \widetilde{q_2}), ..., (n, \widetilde{q_n}).$$

Näitame, et $S - D \geq 0$. Arvestades seost (3.1), saame, et

$$S - D = \sum_{i=1}^{n}(n+1-2q_i)^2 - \sum_{i=1}^{n}(n+1-q_i-\widetilde{q_i})^2 =$$

$$n(n+1)^2 - 4(n+1)\sum_{i=1}^{n} q_i + 4\sum_{i=1}^{n} q_i^2 - n(n+1)^2 +$$

$$+2(n+1)\sum_{i=1}^{n}(q_i+\widetilde{q_i}) - \sum_{i=1}^{n}(q_i+\widetilde{q_i})^2.$$

Kuna summad $\sum_{i=1}^{n} q_i$ ja $\sum_{i=1}^{n} \widetilde{q_i}$ erinevad teineteisest ainult liidetavate järjekorra poolest, siis on nad võrdsed. Seda arvestades saame, et

$$S - D = 4\sum_{i=1}^{n} q_i^2 - \sum_{i=1}^{n}(q_i+\widetilde{q_i})^2 = 2\sum_{i=1}^{n} q_i^2 - 2\sum_{i=1}^{n} q_i\widetilde{q_1} =$$

$$= \sum_{i=1}^{n} q_i^2 - 2\sum_{i=1}^{n} q_i\widetilde{q_i} + \sum_{i=1}^{n} \widetilde{q_i}^2 = \sum_{i=1}^{n}(q_i-\widetilde{q_i})^2 \geq 0.$$

Saime, et suurus S on suuruse D maksimaalne väärtus.

□

Lauset 3.1 arvestades saame Spearmanni astakkorrelatsioonikordaja avaldada kui

$$\rho(X,Y) = 1 - \frac{2D}{S} = 1 - \frac{6D}{n(n^2-1)}.$$

See kordaja kannab nime Spearmanni ρ. See astakkorrelatsioonikordaja iseloomustab monotoonse seose tugevust tunnuste X ja Y vahel. Kui seos on rangelt monotoonselt kasvav, siis $\rho(X,Y) = 1$, rangelt monotoonselt kahanemise korral $\rho(X,Y) = -1$. Näiteks

$$Y = \exp(-X), \quad X \geq 0$$

puhul on $\rho(X,Y) = -1$.

Kuidas aga testida Spearmanni ρ olulisust? Selleks püstitame hüpoteeside paari

$$\begin{cases} H_0 : \rho = 0, \\ H_1 : \rho \neq 0. \end{cases}$$

Osutub, et neid hüpoteese saab testida statistikuga

$$T = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{n-2} \sim t(n-2).$$

Seega saab Spearmanni ρ olulisust kontrollida Studenti t-testiga.

Vaatame, kuidas leida Spearmanni ρ kordajat tarkvara MS Excel abil. Selles tarkvaras leiab astaku funktsioon RANK. Olgu r_i funktsioon, mis seab juhusliku suuruse X väärtusele x_i vastavusse tema astaku. Siis MS Exceli keskkonnas

$$r_i = \mathrm{RANK}(\mathrm{ai}; \mathrm{a1{:}a9}; 1).$$

Selle käsuga saame lahtris ai oleva suuruse astaku lahtrite a1–a9 väärtuste seas. Argument 1 funktsioonis RANK näitab kasvavat järjekorda. Seega

antud juhul

$$\rho(X,Y) = \frac{6\sum_{i=1}^{9}(r_i - q_i)^2}{9(81-1)},$$

kus q_i tähistab juhusliku suuruse Y väärtusele y_i vastavat astakut.

Näide 3.5. Olgu meil 6 vaatlusest koosnev andmestik

X	2	5	7	10	9	8
Y	4	1	9	10	8	6

See andmestik võib näiteks sisaldada 2 testi tulemusi 10-punktisüsteemis. Nendele tulemustele vastavad astakud

r_i	1	2	3	6	5	4
q_i	2	1	5	6	4	3

Leiame suuruse

$$D = \sum_{i=1}^{6}(r_i - q_i)^2 = 8$$

ning sellele vastava Spearmanni astakkorrelatsioonikordaja

$$\rho(X,Y) = 1 - \frac{6D}{n(n^2-1)} = 1 - \frac{48}{6 \cdot (36-1)} \approx 0.771.$$

Seega on kahe testi tulemuste vahel küllaltki tugev monotoonselt kasvav seos. Testime selle seose olulisust. Saame, et teststatistiku T väärtus

$$t \approx \frac{0.771}{\sqrt{1-0.771^2}}\sqrt{4} \approx 2.42.$$

Arvestades kahepoolset hüpoteesi, saame, et väärtusele t vastav olulisustõenäosus *p-value* ≈ 0.072. Kui olulisuse nivooks võtta 0.05, siis peame jääma tõsiasja juurde, et leitud Spearmanni astakkorrelatsioonikordaja ei ole oluline.

3.3.2. Kendalli kordaja

Kendalli astakkorrelatsioonikordaja on veelgi robustsem kui Spearmanni kordaja. Kui Spearmanni astakkorrelatsioonikordaja sisaldab teatud määral infot juhuslike suuruste vahelise seose olemusest (ruutseos, eksponentsiaalne, logaritmiline vms), siis Kendalli oma mitte. Tähistagem juhuslike suuruste X ja Y vahelist Kendalli seosekordajat kui $\tau(X,Y)$. See kordaja saadakse järgmiselt. Valime vaatluste (x_1,y_1), (x_2,y_2), ... ja (x_n,y_n) seast kõikvõimalikud paarid, mida on kokku C_2^n. Nimetame kooskõlalisteks (ingl *concordance*) sellised paarid (x_i,y_i) ja (x_j,y_j), mille puhul

$$(x_i-x_j)(y_i-y_j)>0$$

ning ebakõlalisteks (ingl *discordance*) sellised paarid (x_i,y_i) ja (x_j,y_j), mille puhul

$$(x_i-x_j)(y_i-y_j)<0.$$

Seega vastab kooskõlalistele paaridele tõusev sirge, ebakõlalistele aga langev sirge.

Näide 3.6. Olgu $n=5$. Siis saab kooskõlaliste ja ebakõlaliste paaride hulga leida järgmise tabeli abil:

$(x_2-x_1)(y_2-y_1)$			
$(x_3-x_1)(y_3-y_1)$	$(x_3-x_2)(y_3-y_2)$		
$(x_4-x_1)(y_4-y_1)$	$(x_4-x_2)(y_4-y_2)$	$(x_4-x_3)(y_4-y_3)$	
$(x_5-x_1)(y_5-y_1)$	$(x_5-x_2)(y_5-y_2)$	$(x_5-x_3)(y_5-y_3)$	$(x_5-x_4)(y_5-y_4)$

Tabelit on hõlbus koostada MS Exceli keskkonnas. Selle tabeli positiivsete väärtustega lahtrid vastavad kooskõlalistele, negatiivsete väärtustega lahtrid aga ebakõlalistele paaridele.

Tähistades kooskõlaliste paaride hulga kui c ja ebakõlaliste paaride hulga kui d, saame Kendalli astakkorrelatsioonikordajaks

$$\tau(X,Y)=\frac{c-d}{C_n^2}=\frac{2(c-d)}{n(n-1)}.$$

Saadud seosekordaja kannab nime Kendalli τ.

Vaatame, kuidas testida Kendalli τ olulisust. Osutub, et seda saab teha normaalse aproksimatsiooni abil. Koostame hüpoteeside paari

$$\begin{cases} H_0 : \tau = 0, \\ H_1 : \tau \neq 0. \end{cases}$$

Seda statsitilist hüpoteesi saab kontrollida juhusliku suuruse

$$S = c - d = \sum_{i=1}^{n-1} \sum_{j>i}^{n} \text{sign}(X_j - X_i)\text{sign}(Y_j - Y_i)$$

abil. Selle juhusliku suuruse keskväärtus ja dispersioon on järgmised:

$$E(S) = 0 \text{ ning } D(S) = \frac{n(n-1)(2n+5)}{18}.$$

Seega saame koostada statistiku

$$Z = \frac{3\sqrt{2}(c-d)}{\sqrt{n(n-1)(2n+5)}},$$

mis allub ligikaudselt standardsele normaaljaotusele. Statistikuga Z saab testida Kendalli τ olulisust.

Näide 3.7. Leiame näites 3.5 toodud andmestikule vastava Kendalli τ ning testime selle olulisust. Antud tunnuste puhul saame kooskõlaliste paaride hulgaks $c = 12$ ning ebakõlaliste paaride hulgaks $d = 3$. Seega Kendalli astakkorrelatsioonikordaja

$$\tau(X, Y) = \frac{2 \cdot (12 - 3)}{6 \cdot (6 - 1)} = 0.6.$$

Sellele Kendalli tau väärtusele vastab teststatistiku Z väärtus

$$z = \frac{3 \cdot \sqrt{2} \cdot (12 - 3)}{\sqrt{6 \cdot (6 - 1) \cdot (12 + 5)}} \approx 1.69$$

ning olulisustõenäosus *p-value* ≈ 0.091. Kui olulisuse nivooks võtta 0.05, siis tuleb jääda hüpoteesi $\tau = 0$ juurde.

Märkus. Nii Kendalli τ normaalse aproksimatsiooni kui ka Spearmanni ρ Studenti t-test on ligikaudsed statistilised testid.

Tarkvara R abil saab leida Spearmanni ja Kendalli astakkorrelatsioonikordajaid järgmiselt:

```
x=c(2,5,7,10,9,8)
y=c(4,1,9,10,8,6)
cor.test(x,y,alternative="two.sided",method="spearman",
exact=FALSE)
cor.test(x,y,alternative="two.sided",method="kendall",
exact=FALSE).
```

Argumendi väärtuse `exact=FALSE` korral leitakse Spearmanni astakkorrelatsioonikordajale vastav olulisustõenäosus t-testi abil ning Kendalli τ puhul kasutatakse selleks normaalset aproksimatsiooni.

3.4. Mitteparameetrilised statistilised testid

Anname ülevaate statistilises andmeanalüüsis enim kasutatavatest mitteparameetrilistest testidest. Nendes kasutatakse uuritava suuruse väärtuse asemel tema astakuid.

3.4.1. Wilcoxoni astakute test

Olgu meil kaks valimit (kas sõltumatut või siis kordusmõõtmiste omad), mida tahame võrrelda. Kui mõlemad valimid alluksid normaaljaotusele, siis sobib selleks eelmises peatükis kirjeldatud t-test. Vaatleme nüüd juhtu, kui Studenti t-testi eeldused pole täidetud. Näiteks olukorda, kui mõlema valimi jaotused on kaugel normaaljaotustest. Siis võib populatsioonide võrdlusteks kasutada Wilcoxoni testi, mis põhineb astakute summadel. Sageli nimetatakse seda testi ka Wilcoxon-Mann-Whitney testiks.

Olgu esimeses valimis n_1 ja teises valimis n_2 elementi. Me leiame mõlemale populatsioonile astakute summad, milleks olgu vastavalt T_1 ja T_2.

Olgu $n = n_1 + n_2$. Siis olenemata valimi mahtudest n_1 ja n_2

$$T_1 + T_2 = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Seega, kui T_1 väärtus suureneb, siis T_2 oma väheneb ning vastupidi.

Demonstreerime astakute testi ühe konkreetse näitega.

Näide 3.8. Võrreldi ühe tekstitöötluse tarkvara kasutuse oskust tavakasutajate ja kutseliste programmeerijate vahel. Taheti tõestada nende erinevust. Oskust hinnati punktidega skaalal 1 kuni 100. Tulemused on esitatud tabelis 3.2.

Tabel 3.2. Tavakasutaja ja programmeerija oskuste hinded ning nende astakud

Tavakasutaja oskused	Astakud	Programmeerija oskused	Astakud
35	5	45	7
50	8	60	10
25	3	40	6
55	9	90	13
10	1	65	11
30	4	85	12
20	2	95	14

Tabeli 3.2 põhjal järeldub, et mõlema grupi valimi mahtudeks on 7 ning esimese valimi astakute summa $T_1 = 32$ ja teise valimi astakute summa $T_2 = 73$. Olgu D_1 tavakasutajate oskuspunktide jaotus ning D_2 programmeerijate oskuspunktide jaotus.

Olgu testi olulisuse nivoo $\beta = 0.05$. Võtame teststatistikuks astakute summa T_1 (kuid sama hästi võime selleks võtta ka T_2). Võrdleme antud valimi põhjal saadud statistiku T_1 väärtust tabelis 3.3 toodud astakute summa alumise usalduspiiri T_L ja ülemise usalduspiiri T_U väärtustega.

Hüpoteeside paari sõnastame järgmiselt:

$$\begin{cases} H_0 : D_1 \text{ ja } D_2 \text{ on võrdsed,} \\ H_1 : D_1 \text{ ja } D_2 \text{ on erinevad.} \end{cases}$$

Tabel 3.3. Wilcoxoni testi ülemised ja alumised kriitilised väärtused olulisuse nivool 0.025 ühepoolse ja olulisuse nivool 0.05 kahepoolse testi korral

n_2 \ n_1	3	3	4	4	5	5	6	6	7	7	8	8	9	9
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65
7	7	26	13	35	20	45	28	56	**37**	**68**	39	73	41	78
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114

Tablist 3.3 järeldub, et nullhüpoteesi kummutamiseks peab $T_1 \leq 37$ või $T_1 \geq 68$. Meie valimi puhul $T_1 = 32$. Seega võime väita, et tavakasutaja oskused antud tarkvaraga töötamisel on oluliselt väiksemad kui programmeerija omad.

Uurime nüüd, kuidas saadi tabelisse 3.3 vastavad arvud. Vaatame juhtu, kus $n_1 = n_2 = 4$. Meil tuleb selline väärtus T_L, et

$$P(T_1 \leq T_L) = P(T_1 = 10) + P(T_1 = 11) + ... + P(T_1 = T_L) \approx 0.025.$$

Kuna valimisse kaasamise tõenäosused on kõikidel elementidel võrdsed, siis iga permutatsiooni P_i realisatsiooni tõenäosus

$$P(P_i) = \frac{1}{8!},$$

kus $i = 1, 2, ..., 8!$. Antud juhul on vähim astakute summa $T_1 = 1 + 2 + 3 + 4 = 10$. Võimalusi, et $T_1 = 10$ on antud juhul 4!4!. Seega saame, et

$$P(T_1 = 10) = \frac{4!4!}{8!} = \frac{1}{70} = 0.0143.$$

Vaatame juhtu, kui $T_1 = 11$. Sellised juhul saavad olla esimese valimi astakud 1, 2, 3 ja 5. Seega

$$P(T = 11) = \frac{4!4!}{8!} = \frac{1}{70} = 0.0143$$

ning

$$P(T_1 \leq 11) = (T_1 = 10) + P(T_1 = 11) = 0.0143 + 0.0143 = 0.0286.$$

Järelikult $P(T_1 \leq T_L) \approx 0.025$, kui $T_L = 11$.

Kui valimi maht on piisavalt suur (enamasti juhul, kui $n_1 > 10$ ja $n_2 > 10$), siis saame Wilcoxoni testi puhul rakendada tsentraalset piirteoreemi. Sellisel juhul on statistik

$$Z = \frac{T_1 - E(T_1)}{\sqrt{D(X_1)}}$$

ligikaudu standardse normaaljaotusega. Leiame keskväärtuse $E(T_1)$ ning dispersiooni $D(T_1)$.

Lause 3.2. Olgu esimese populatsiooni valimi maht n_1 ning selle populatsiooni elementide astakute summa T_1. Siis

$$E(T_1) = \frac{n_1(n+1)}{2} \text{ ja } D(T_1) = \frac{n_1(n-n_1)(n+1)}{12}.$$

Tõestus. Olgu esimese populatsiooni i-nda elemendi astak juhuslik suurus

$$S_i = \{1, 2, ..., n\}, \quad P(S_i = i) = \frac{1}{n}, \quad i = 1, 2, ..., n.$$

Siis $T_1 = \sum_{i=1}^{n_1} S_i$ ning tema keskväärtus

$$E(T_1) = E\left(\sum_{i=1}^{n_1} S_i\right) = \sum_{i=1}^{n_1} E(S_i) = \sum_{i=1}^{n_1} \frac{1}{n} \sum_{j=1}^{n} j =$$

$$= \sum_{i=1}^{n_1} \frac{1}{n} \frac{n(n+1)}{2} = \frac{n_1(n+1)}{2}.$$

Astakute summa T_1 dispersioon

$$D(T_1) = D(\sum_{i=1}^{n_1} S_i) = \sum_{i=1}^{n_1} D(S_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n_1} \operatorname{cov}(S_i, S_j).$$

Astaku dispersioon

$$D(S_i) = E(S_i^2) - E^2(S_i) = \frac{1}{n}\sum_{i=1}^{n} i^2 - \frac{(n+1)^2}{4} =$$

$$= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{2(n+1)(2n+1) - 3(n+1)^2}{12} =$$

$$= \frac{(n+1)(4n+2-3n-3)}{12} = \frac{n^2-1}{12}.$$

Astakute S_i ja S_j vaheline kovariatsioon

$$\mathrm{cov}(S_i, S_j) = E(S_iS_j) - E(S_i)E(S_j) = E(S_iS_j) - \frac{(n+1)^2}{4}.$$

Kogu raskus seisneb keskväärtuse $E(S_iS_j)$ leidmises. Tõenäosus

$$P(S_i = i, S_j = j) = \frac{1}{n(n-1)}, \quad i, j = 1, 2, ..., \quad i \neq j.$$

Seega

$$E(S_iS_j) = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n} ij = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} ij =$$

$$= \frac{2}{n(n-1)}\sum_{i=1}^{n-1} i\frac{(1+n+i)(n-i)}{2} = \frac{1}{n(n-1)}\sum_{i=1}^{n-1} i(n-i+n^2-i^2) =$$

$$= \frac{1}{n(n-1)}\left(n\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} i^2 + n^2\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} i^3\right).$$

Summa

$$\sum_{i=1}^{n-1} i^3 = \left(\sum_{i=1}^{n-1} i\right)^2 = \frac{n^2(n-1)^2}{4}.$$

Järelikult

$$E(S_iS_j) = \frac{1}{n(n-1)}\times$$

$$\times\left(\frac{n^2(n-1)}{2} - \frac{n(n-1)(2n-1)}{6} + \frac{n^2(n-1)n}{2} - \frac{n^2(n-1)^2}{4}\right) =$$

$$= \frac{6n - 4n + 2 + 6n^2 - 3n^2 + 3n}{12} = \frac{3n^2 + 5n + 2}{12} = \frac{(n+1)(3n+2)}{12}.$$

Seega saame kovariatsiooniks

$$\operatorname{cov}(S_i, S_j) = \frac{(n+1)(3n+2)}{12} - \frac{(n+1)^2}{4} =$$

$$= \frac{(n+1)(3n+2-3n-3)}{12} = -\frac{n+1}{12}.$$

Leiame nüüd astakute summa dispersiooni arvestades, et $\operatorname{cov}(S_i, S_j) = \operatorname{cov}(S_j, S_i)$. Saame, et

$$D(T_1) = \frac{n_1(n^2-1)}{12} - 2C_{n_1}^2 \frac{n+1}{12} = \frac{n_1(n^2-1)}{12} - 2\frac{n_1!}{2!(n_1-2)!}\frac{n+1}{12} =$$

$$= \frac{n_1(n^2-1) - n_1(n_1-1)(n+1)}{12} = \frac{n_1(n+1)(n-1-n_1+1)}{12} =$$

$$= \frac{n_1(n-n_1)(n+1)}{12}.$$

□

Lausest 3.2 saame teha ühe huvitava järelduse.

Järeldus 3.1. Astakutevaheline lineaarne korrelatsioonikordaja

$$\operatorname{corr}(S_i, S_j) = -\frac{1}{n-1}.$$

Tõestus. Tõepoolest

$$\operatorname{corr}(S_i, S_j) = \frac{\operatorname{cov}(S_i, S_j)}{\sqrt{S_i}\sqrt{S_j}} = -\frac{\frac{n+1}{12}}{\sqrt{\frac{n^2-1}{12}}\sqrt{\frac{n^2-1}{12}}} =$$

$$= -\frac{n+1}{n^2-1} = -\frac{1}{n-1}.$$

□

Kokkuvõttes saime, et statistik

$$Z = \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1(n-n_1)(n+1)}{12}}}.$$

Kui laadida tarkvarale R pakett `stats`, siis saab selles Wilcoxoni testi läbi viia käsuga

```
wilcox.test(x, y,
alternative = c("two.sided", "less", "greater"),
```

mille argumendid `x` ja `y` määravad uuritavad tunnused ning argument `alternative = c("two.sided", "less", "greater")` näitab, kas on tegemist kahepoolse, vasak- või parempoolse testiga.

3.4.2. Kvantiili test

Eelmistes peatükkides käsitlesime statistilisi teste, kus testisime uuritava suuruse keskväärtust. Sageli aga huvitavad meid just n-ö ääreväärtused, näiteks 0.025- ja 0.975-kvantiilid. Vaatleme olukorda, kus tahame mõõtmistulemuste $(x_1, x_2, ..., x_n)$ abil testida uuritava suuruse α-kvantiili. Olgu meil hüpoteeside paar

$$\begin{cases} H_0 : x_\alpha = x_0, \\ H_1 : x_\alpha \neq x_0. \end{cases}$$

Teeme mõõtmistulemuste hulgast m-elemendilise juhusliku valimi, kus valik on tagasipanekuga. Olgu statistik S elementide hulk, mis ei ületa väärtust x_0. Kuna valik oli sõltumatu (s.t tagasipanekuga), siis juhuslik suurus S allub binoomjaotusele parameetritega m ja α. Seega keskväärtus $E(S) = m\alpha$ ning standardhälve $\sigma_S = \sqrt{m\alpha(1-\alpha)}$. Tsentraalse piirteoreemi põhjal on statistik

$$Z = \frac{S - m\alpha}{\sqrt{m\alpha(1-\alpha)}}$$

ligikaudu standardse normaaljaotusega.

Näide 3.9. Käesoleva näitega anname ühe võimaluse realiseerimaks kvantiili testi tarkvaras R. Olgu meil valimi $\mathbf{X}$ realisatsioon $\mathbf{x} = (2, 1, 7, 4, 8, 9)^\intercal$. Selle realisatsiooni põhjal kontrollitakse järgmist hüpoteeside paari 0.8-kvantiilile:

$$\begin{cases} H_0 : x_{0.8} = 8, \\ H_1 : x_{0.8} \neq 8. \end{cases}$$

Viime selle testi läbi 300 korda. Iga kord olgu $m = 1000$. Seda saab teha tarkvara R järgmise programmiga:

```
x=c(2,1,7,4,8,9)
y=0
Pvalue=0
for (j in 1:300){
for (i in 1:1000) y[i]=sample(x,1)
S=length(sample(y[y<=8]))
Z=(S-0.8*1000)/sqrt(0.8*0.2*1000)
Pvalue[j]=2*(1-pnorm(abs(Z)))}
Pvalue.
```

Vundiks saame 300 erinevat olulisustõenäosust, mille põhjal leitud vahemikhinnangu alusel teeme otsustuse.

3.4.3. Kruskal-Wallise H-test

Selle testiga võrreldakse enam kui kahte sõltumatut valimit. Tegemist on dispersioonanalüüsi (ANOVA) üldistamisega juhule, mil normaaljaotuse eeldus pole täidetud. Antud juhul omistatakse esmalt igale vaatlusele astak. Seejärel leiatakse Kruskal-Wallise H-statistik

$$H = \frac{12}{n(n-1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1), \tag{3.2}$$

kus k tähistab valimite hulka, T_i on i-nda valimi astakute summa ning n_i tähistab vaatluste hulka selles valimis. Statistik (3.2) allub ligikaudu χ^2-jaotusele vabadusastmete arvuga $k-1$. Kui klassikalise ANOVA puhul võrdlesime erinevate populatsioonide keskväärtuseid, siis Kruskal-Wallise

H-testi puhul võrdleme mediaane. Olgu $\theta_1, \theta_2, ..., \theta_k$ erinevate valimite mediaanid. Siis saame formuleerida hüpoteeside paari järgmiselt:

$$\begin{cases} H_0 : \theta_1 = \theta_2 = ... = \theta_k, \\ H_1 : \text{vähemalt 2 populatsiooni astakute vahel on oluline erinevus.} \end{cases}$$

Kui valimite elementidel esineb võrdseid astakuid, siis tuleb statistikut (3.2) korrigeerida parandiga

$$C_H = 1 - \frac{\sum_{j=1}^{l}(R^3 - R)}{n^3 - n},$$

kus l on sarnaste väärtustega astakute hulk ning R tähistab sarnaste väärtustega astakute arvu vastavas hulgas. Parandatud statistik $H_C = \frac{H}{C_H}$. Kui $R = l = 1$, siis $C_H = H$.

Demonstreerime Kruskal-Wallise H-testi konkreetse näitega.

Näide 3.10. Uuriti täiskasvanute sotsiaalse suhtlemise aktiivsuse ja enesekindluse vahelist sõltuvust. Selleks jagati 17 uuringus osalejat suhtlemise taseme poolest 3 gruppi: kõrge, keskmine ja madal. Kõik osalejad tegid läbi testi, mille põhjal hinnati nende enesekindlust 25 punkti skaalal. Saadi järgmised tulemused:

Kõrge	Keskmine	Madal
21	19	7
23	5	8
18	10	15
12	11	3
19	9	6
20		4

Kui igale testi tulemusele seati vastavusse astak, siis võttis tabel järgmise kuju:

16	13.5	5
17	3	6
12	8	11
10	9	1
13.5	7	4
15		2

Järgnevalt leiti igale suhtlustaseme grupile astakute summad T_i, $i = 1, 2, 3$.

Kõrgema taseme puhul saadi

$$T_1 = 16 + 17 + 12 + 10 + 13.5 + 15 = 83.5, \quad n_1 = 6.$$

Keskmise taseme puhul

$$T_2 = 13.5 + 3 + 8 + 9 + 7 = 40.5, \quad n_2 = 5.$$

Madala taseme puhul

$$T_3 = 5 + 6 + 11 + 1 + 4 + 2 = 29, \quad n_3 = 6.$$

Seose (3.2) põhjal saadi statistiku H väärtuseks

$$h = \frac{12}{17(17+1)}\left(\frac{83.5^2}{6} + \frac{40.5^2}{5} + \frac{29^2}{6}\right) - 3(17+1) = 9.93.$$

Antud juhul esines võrdseid astakuid ühel korral ning neid astakuid oli 2 (väärtus 13.5). Seega parand

$$c_h = 1 - \frac{2^3 - 2}{17^3 - 17} = 0.9988.$$

Niisiis saame teststatistiku lõplikuks väärtuseks

$$h_c = \frac{h}{c_h} = \frac{9.93}{0.9988} \approx 9.94.$$

Antud juhul $H \sim \chi^2(2)$. Seega olulisustõenäosus

$$p\text{-}value = 2 \cdot \min\{P(H > 9.94); P(H \leq 9.94)\} \approx 0.014.$$

Kui olulisuse nivoo $\beta = 0.05$, siis võib lugeda tõestatuks väite, et enesekindlus sõltub suhtlemisaktiivsusest.

Tarkvaras R saab Kruskal-Wallise H-testi läbi viia käsuga

`kruskal.test(Hinne~Tase, data = x)`.

Näites 3.10 olev andmestik, mis on tähistatud kui `x`, tuleb algselt sisestada objekt-tunnus maatriksi kujul. Ühes veerus on enesekindluse hinnang, teises suhtlemistaseme grupi näitaja.

3.4.4. Friedmanni test

See mitteparameetriline test on oma nime saanud selle väljaarendaja Nobeli preemia laureaadi Milton Friedmanni (1912–2006) järgi. Tegu on mitteparameetrilise ANOVA testiga kordusmõõtmiste puhul. See tähendab, et mõõtmised on tehtud samadel ojektidel, kuid erinevatel aegadel või erinevates tingimustes. Olgu erinevate mõõtmisgruppide astakute summad $T_1, T_2, ..., T_k$ ning olgu m iga grupi valimi maht. Siis statistik

$$F_r = \frac{12}{mk(k+1)} \sum_{i=1}^{k} T_i^2 - 3m(k+1).$$

Statistiku F_r jaotus on ligilähedane χ^2-jaotusele vabadusastemete arvuga $k-1$. See lähend hakkab toimima, kui mõõtmisgruppide hulk k ning iga grupi valimi maht m on suuremad kui 5.

Selgitame Friedmanni testi olemust järgmise näitega.

Näide 3.11. Näide puudutab erinevate metallide korrosiooni uurimist. Uuriti 3 erinevat laeva pärast kuuajalist viibimist erinevates kliimatingimustes. Mõõdeti nende 3 laeva metallkonstruktsioonis oleva 10 metallitüübi (1, 2, ..., 10) korrosiooni astet (%) peale reise. Iga laeva puhul vaadeldi samu metallitüüpe. Sooviti kontrollida hüpoteeside paari:

$$\begin{cases} H_0 : \text{korrosiooniastme jaotus on kõigi 3 laeva puhul sama,} \\ H_1 : \text{vähemalt 1 laeva korrosiooniastme jaotus erineb teistest.} \end{cases}$$

Mõõtmistulemuste põhjal koostati korrosiooniastmetest järgmine tabel:

Metall	Laev 1	Astakud	Laev 2	Astakud	Laev 3	Astakud
1	21	2	23	3	15	1
2	29	2	30	3	21	1
3	16	1	19	3	18	2
4	20	3	19	2	18	1
5	13	2	10	1	14	3
6	5	1	12	3	6	2
7	18	2.5	18	2.5	12	1
8	26	2	32	3	21	1
9	17	2	20	3	9	1
10	4	2	10	3	2	1

Antud juhul $k = 3$ ning $m = 10$. Astakute summadeks saadi tabeli põhjal järgmised tulemused: laeva 1 puhul $T_1 = 19.5$, laeva 2 puhul $T_2 = 26.5$ ning laeva 3 puhul $T_3 = 14$. Mõõtmistulemuste põhjal leiti statistiku F_r väärtus

$$f_r = \frac{12}{10 \cdot 3 \cdot (3+1)}(19.5^2 + 26.5^2 + 14^2) - 3 \cdot 10 \cdot (3+1) = 7.85.$$

Antud juhul on statistik F_r jaotuseks χ^2-jaotus vabadusastmete arvuga $k - 1 = 2$. Tarkvara MS Exceli abil saadi olulisustõenäosuseks

$$p\text{-}value = P(f_r > F_r) = 1 - \text{CHISQ.DIST}(7.85; 2; \text{TRUE}) \approx 0.02.$$

Kui olulisuse nivooks võtta 0.05, siis võib saadud mõõtmistulemuste põhjal nullhüpoteesi H_0 kummutada ehk lugeda tõestatuks sisukas hüpotees H_1.

Viimaks Friedmanni testi läbi tarkvaras R tuleb esmalt alla laadida pakett `stats`. Seejärel tuleb sisestada käsk

`friedman.test(y~x1|x2,data=x)`,

kus `x` tähistab sisse loetud andmestikku, `y` uuritavat tunnust ning `x1` ja `x2` faktortunnuseid. Andmed `x` tuleb eelnevalt esitada objekt-tunnus maatriksina, milles on 30 rida (3 laeva, iga laeva konstruktsioonis 10 erinevat metallitüüpi) ning 3 veergu (tunnused korrosiooni aste, laev ja metall). Näites 3.11 on tunnuseks `y` korrosiooniaste, tunnuseks `x1` laev (1, 2 või 3) ja tunnuseks `x2` metalli tüüp (1, 2,... või 10). Tähistusega `|x2` näidatakse, et iga laeva puhul on uuritud erinevaid metallitüüpe.

3.4.5. Kolmogorov-Smirnovi test

Vaatame lähemalt statistilist testi, mida tuntakse kui Kolmogorov-Smirnovi testi. Kolmogorov-Smirnovi statistilist testi kasutatakse mitteparameetrilises statistikas kllaltki palju. Nimelt testitakse sellega jaotuse sobivust. Mitteparameetrilises statistikas aga puudub eelnev eeldus valimi jaotusest.

Käsitleme Kolmogorov-Smirnovi testi kahte juhtu. Esimene neist puudutab kahe empiirilise jaotuse võrdlust. Teine aga empiirilise jaotuse võrdlust teoreetilisega.

Kahe valimi jaotuste võrdlemine

Testime kaht erinevat valimit järgmistel eeldustel:

1) olgu meil valimid $\mathbf{X} = (X_1, X_2, ..., X_{n_1})^\intercal$ ning $\mathbf{Y} = (Y_1, Y_2, ..., Y_{n_2})^\intercal$, mille elemendid on sõltumatud ja sama jaotusega (s.s.j);

2) need kaks valimit on sõltumatud.

Koostame järgmise hüpoteeside paari:

$$\begin{cases} H_0 : \text{valimid } \mathbf{X} \text{ ja } \mathbf{Y} \text{ on sama jaotusega,} \\ H_1 : \text{valimid } \mathbf{X} \text{ ja } \mathbf{Y} \text{ on eri jaotustega.} \end{cases}$$

Selle kontrollimiseks koostame statistiku Z, mida nimetatakse Kolmogorov-Smirnovi statistikuks. Esimese sammuna leiame valimite $\mathbf{X}$ ja $\mathbf{Y}$ empiirilised jaotusfunktsioonid $F_{n_1}(t)$ ja $G_{n_2}(t)$. Iga reaalarvulise t korral olgu

$$F_{n_1}(t) = \frac{\mid \{\text{vaatlused } x_i \mid x_i \leq t\} \mid}{n_1}$$

ning

$$G_{n_2}(t) = \frac{\mid \{\text{vaatlused } y_j \mid y_j \leq t\} \mid}{n_2}.$$

Seejärel leiame igale vaatlusele vastava empiiriliste jaotuste erinevuse

$$D = \mid F_{n_1}(t) - G_{n_2}(t) \mid .$$

Lõpuks leiame Kolmogorov-Smirnovi statistiku

$$Z = D_{\max}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \tag{3.3}$$

kus $D_{\max}$ on empiiriliste jaotuste maksimaalne erinevus ehk suuruse D maksimaalne väärtus. Publikatsioonis [38] on toodud eeskiri leidmaks statistiku Z väärtusele vastavat olulisustõenäosust (*p-value*). Seda võib esitada järgmiselt:

1) kui $0 \leq Z < 0.27$, siis $p\text{-}value = 1$;

2) kui $0.27 \leq Z < 1$, siis

$$p\text{-}value = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25}), \tag{3.4}$$

kus

$$Q = \exp(-1.233701Z^{-2}); \tag{3.5}$$

3) kui $1 \leq Z < 3.1$, siis

$$p\text{-}value = 2(Q - Q^4 + Q^9 - Q^{16}), \tag{3.6}$$

kus

$$Q = \exp(-2Z^2); \tag{3.7}$$

4) kui $Z \geq 3.1$, siis $p\text{-}value = 0$.

Demonstreerime statistiku Z leidmist ning selle väärtuse põhjal otsustuse tegemist ühe konkreetse näitega.

Näide 3.12. Võrreldakse kahte võõrkeele õpetamise meetodit. Võõrkeele kiirkursus viidi läbi kahel erineval meetodil: meetodil A ja meetodil B. Pärast kursust tehti test ning tulemusi hinnati 100 punkti skaalal. Saadi järgmised punktisummad:

Meetod A	Meetod B
48	14
40	18
39	20
50	10
41	12
38	100
53	17

Võrdlemaks meetodite erinevusi püstitati hüpoteesid:

$$\begin{cases} H_0 : \text{meetodite A ja B punktisumma jaotused on samad,} \\ H_1 : \text{meetodite A ja B punktisumma jaotused on erinevad.} \end{cases}$$

Selle hüpoteeside paari kontrolliks koostati punktisummade põhjal tabel, milles kajastuvad meetoditele A ja B vastavate punktisummade jaotused:

Punktisumma z_i	$F_7(z_i)$	$G_7(z_i)$	$\mid F_7(z_i) - G_7(z_i) \mid$
10	0	$\frac{1}{7}$	$\frac{1}{7}$
12	0	$\frac{2}{7}$	$\frac{2}{7}$
14	0	$\frac{3}{7}$	$\frac{3}{7}$
17	0	$\frac{4}{7}$	$\frac{4}{7}$
18	0	$\frac{5}{7}$	$\frac{5}{7}$
20	0	$\frac{6}{7}$	$\frac{6}{7}$
38	$\frac{1}{7}$	$\frac{6}{7}$	$\frac{5}{7}$
39	$\frac{2}{7}$	$\frac{6}{7}$	$\frac{4}{7}$
40	$\frac{3}{7}$	$\frac{6}{7}$	$\frac{3}{7}$
41	$\frac{4}{7}$	$\frac{6}{7}$	$\frac{2}{7}$
48	$\frac{5}{7}$	$\frac{6}{7}$	$\frac{1}{7}$
50	$\frac{6}{7}$	$\frac{6}{7}$	0
53	1	$\frac{6}{7}$	$\frac{1}{7}$
100	1	1	0

Sellest tabelist järeldub, et empiiriliste jaotuste maksimaalne erinevus $D_{\max} = \frac{6}{7} \approx 0.86$. Seega statistiku Z väärtus

$$z = 0.86\sqrt{\frac{7 \cdot 7}{7+7}} = 0.86 \cdot \sqrt{3.5} \approx 1.604.$$

Antud juhul $1 \leq z < 3.1$. Järelikult tuleb olulisustõenäosuse leidmiseks kasutada seoseid (3.6) ja (3.7). Nende põhjal saame, et teststatistiku Q väärtus

$$q = \exp(-2 \cdot 1.604^2) \approx 0.0058$$

ning sellele vastav olulisustõenäosus

$$p\text{-}value = 2(0.0058 - 0.0058^4 + 0.0058^9 - 0.0058^{16}) \approx 0.012.$$

Kui olulisuse nivooks võtta 0.05, siis saame väita, et meetodi A ning meetodi B tulemuste jaotus oli erinev.

Valimi empiirilise jaotuse võrdlemine teoreetilise jaotusega

Olgu meil valimi $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ empiiriline jaotus $F_n(x)$. Tahame kontrollida, kuivõrd lähedal on empiiriline jaotus teoreetilisele jaotusele $F(x)$, milleks võib olla näiteks normaaljaotus, eksponentjaotus, Poissoni jaotus või muu jaotus. Kontrollimiseks sõnastame valimi esindaja X jaoks hüpoteeside paari:

$$\begin{cases} H_0 : X \sim F, \\ H_1 : X \text{ allub muule jaotusele.} \end{cases}$$

Kontrollimaks seda jagame uuritava suuruse X suurusklassidesse

$$[x_1; x_2), [x_2; x_3), ..., [x_m; x_{m+1})$$

ning leiame igale klassile sageduse vastavalt $n_1, n_2, ..., n_m$. Kusjuures

$$\sum_{i=1}^{m} n_i = n.$$

Olgu $p_i = \dfrac{n_i}{n}$, $i = 1, 2, ..., m$. Siis saame empiirilise ja teoreetilise jaotusfunktsiooni väärtustele järgmise tabeli:

x	$[x_1; x_2)$	$[x_2; x_3)$	$[x_3; x_4)$	$\cdots$	$[x_{m-1}; x_m)$	$[x_m; x_{m+1})$
$F_n(x)$	0	p_1	$p_1 + p_2$	$\cdots$	$p_1 + ...p_{m-1}$	1
$F(x)$	$F(x_1)$	$F(x_2)$	$F(x_3)$	$\cdots$	$F(x_{m-1})$	$F(x_m)$

Leiame Kolmogorovi statistiku

$$D = \max_{1 \le i \le m} \mid F_n(x_i) - F(x_i) \mid . \tag{3.8}$$

Milline on statistiku D jaotus? Vastuseks sellele küsimusele sõnastame järgmise teoreemi.

Teoreem 3.2. Olgu jaotusfunktsioon $F(x)$ pidev ning olgu

$$D_n = \sup \mid F_n(x) - F(x) \mid .$$

Siis iga $\lambda \geq 0$ korral

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq \lambda) = L(\lambda),$$

kus jaotusfunktsioon

$$L(\lambda) = 1 - \sum_{i=-\infty}^{\infty} (-1)^i \exp(-2i^2\lambda^2).$$

Teoreemi 3.2 põhjalik tõestus on esitatud teadusartiklis [11]. Funktsiooni $F_n(x)$ võib käsitleda selles teoreemis kui empiirilist jaotusfunktsiooni.

Olulisuse nivoole β vastav kriitiline väärtus λ_β avaldub kui

$$P(\sqrt{n}D > \lambda_\beta) = \beta.$$

Mõningatele olulisuse nivoodele vastavad kriitilised väärtused on toodud alljärgnevas tabelis.

Tabel 3.4. Kolmogorov-Smirnovi testi kriitilised väärtused erinevate olulisuse nivoode korral

β	0.005	0.01	0.025	0.05	0.1
λ_β	1.73	1.63	1.48	1.36	1.22

Toome näite Kolmogorov-Smirnovi testist kontrollimaks uuritava suuruse allumist normaaljaotusele.

Näide 3.13. Olgu X juulikuine keskmine ööpäevane temperatuur kraadides Celsiuse järgi. Ööpäeva keskmised temperatuurid jagunesid järgmiselt:

Temperatuurid $[x_i; x_{i+1})$	Päevade hulk n_i
$[14; 15)$	3
$[15; 16)$	4
$[16; 17)$	7
$[17; 18)$	8
$[18; 19)$	4
$[19; 20)$	3
$[20; 21)$	1
$[21; 22)$	1

Hüpoteeside paar on antud juhul järgmine:

$$\begin{cases} H_0 : X \sim \mathcal{N}(\mu, \sigma), \\ H_1 : X \text{ allub muule jaotusele.} \end{cases}$$

Leiame koostatud sagedustabeli põhjal hinnangud parameetritele μ ja σ. Keskväärtuse hinnanguks saame, et

$$\overline{x} = \frac{1}{31} \sum_{i=1}^{m} n_i \frac{x_i + x_{i+1}}{2}.$$

Standardhälbe σ nihketa hinnang

$$s = \sqrt{\frac{1}{30} \sum_{i=1}^{m} n_i \left(\frac{x_i + x_{i+1}}{2} - \overline{x} \right)^2}.$$

Kümnendiku täpsusega saame, et $\overline{x} = 17.3$ ja $s = 1.7$. Seega teoreetiline jaotusfunktsioon

$$F(x_i) = 0.5 + \Phi\left(\frac{x_i - \overline{x}}{s} \right).$$

Empiirilise ja teoreetilise jaotuse erinevus on toodud järgmises tabelis:

$[x_i; x_{i+1})$	$F_{31}(x)$	$F(x)$	$\mid F_{31}(x) - F(x) \mid$
$[14; 15)$	0	0.029	0.029
$[15; 16)$	0.226	0.094	0.132
$[16; 17)$	0.452	0.23	0.222
$[17; 18)$	0.71	0.437	0.273
$[18; 19)$	0.839	0.663	0.176
$[19; 20)$	0.935	0.841	0.094
$[20; 21)$	0.968	0.943	0.025
$[21; 22]$	1	0.985	0.015

Seose (3.8) abil saame Kolmogorovi statistikuks $D = 0.273$ ning see vastab keskmisele temperatuurile poolvahemikus $[17, 18)$. Lõpliku otsuse normaaljaotuse kohta teeme suuruse $\sqrt{n}D = \sqrt{310}0.273 \approx 1.52$ põhjal. Tabelist 3.4 järeldub, et leitud väärtusele vastav *p-value* $\in [0.01; 0.025]$. Seega võime olulisuse nivool 0.05 ümber lükata väite, et juulikuu keskmine ööpäevane temperatuur allub normaaljaotusele.

Tarkvara R abil saab teostada Kolmogorov-Smirnovi testi käsuga

`ks.test(x,y)`,

kus `x` ja `y` tähistavad kahte võrreldavat valimit. Võrdlemaks valimi `x` jaotust standardse normaaljaotusega tuleb sisestada käsk

`ks.test(x,"pnorm",0,1)`.

Kui installeerida tarkvarale R pakett `kolmim`, siis saab leida statistikule $\sqrt{n}D$ vastava olulisustõenäosuse käsuga

`pkolm(D,n)`.

3.4.6. Anderson-Darlingi test

Kuigi Kolmogorov-Smirnovi test on olnud laialdaselt kasutusel, on sellel mõningad puudused. Neist kaks peamist on järgmised:

1) suure valimi mahu n korral kummutab see test nullhüpoteesi pea alati;

2) valimi servades ehk suurte ja väikeste väärtuste juures on Kolmogorov-Smirnovi test tundetu.

Teise puuduse vastu on hakatud viimasel ajal kasutama Anderson-Darlingi testi, mis on Kolmogorov-Smirnovi testi modifikatsioon. Anderson-Darlingi test on jaotustevaba, see tähendab, et testitavas jaotuses ei ole vaja määrata parameetreid. Test avaldati 1954. aastal teadusartiklis [1]. See põhineb juhuslikul suurusel

$$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} f(x)dx. \tag{3.9}$$

Seose (3.9) lugejas on empiirilise ja teoreetilise jaotuse vahe ruut. Seega, mida suurem on statistiku A väärtus, seda rohkem on põhjust kahelda valimi allumises jaotusele F. Publikatsioonis [1] näidati, et statistik

$$A = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \{\ln(F(X_{(i)})) - \ln(F(X_{(n+1-i)}))\},$$

kus $X_{(1)} < X_{(2)} < ... < X_{(i)} < ... < X_{(n)}$ on valimist $\mathbf{X}$ saadud järkstatistikud.

Kontrollimaks tarkvara R abil valimi allumist normaaljaotusele esmalt installeerida käsuga

```
install.packages("nortest")
```

pakett `nortest`. Seejärel on võimalik läbi viia Anderson-Darlingi test käsuga

```
ad.test(x)
```

kus `x` tähistab valimi realisatsiooni.

3.4.7. Wilk-Šapiro test

Tutvume statistilise testiga, mis testib, kas valim $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$ esindab normaaljaotusega üldkogumit. Testi idee ja selle realiseermise publitseerisid esmakordselt Samuel Šapiro ning Martin Wilk 1965. aastal teadusartiklis [40]. Nullhüpotees väidab, et valim esindab normaaljaotusega populatsiooni. Selle kummutamiseks koostatakse teststatistik

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{x})^2},$$

kus $X_{(i)}$ tähistab i-ndat järkstatistikut, $\overline{x} = \dfrac{X_1 + X_2 + ... + X_n}{n}$ on valimi põhjal leitud aritmeetiline keskmine ning vektor

$$(a_1, a_2, ..., a_i, ..., a_n)^\intercal = \frac{\mu^\intercal \mathbf{C}^{-1}}{(\mu^\intercal \mathbf{C}^{-1}\mathbf{C}^{-1}\mu)^{\frac{1}{2}}},$$

milles $\mu = (\mu_1, \mu_2, ..., \mu_i, ..., \mu_n)^\intercal$ on standardse normaaljaotusega järkstatistikute keskväärtuste vektor ning $\mathbf{C}$ nendevaheline kovariatsioonimaatriks.

Viime Wilk-Šapiro testi läbi tarkvara R abil. Selleks tuleb kasutada R-i paketti `stats`. Vaatame järgmist juhtu:

```
x=cbind(11.6,12.9,8.5,14.1,7.9,9.8,13.1,15.5,8.4)
shapiro.test(x)
```

Antud juhul on tegemist valimiga, mille maht on 9. Selle normaaljaotusele vastamist testib käsk `shapiro.test(x)`, mis annab antud juhul väljundiks

```
W = 0.9227,  p-value = 0.4151.
```

Teststatistiku W väärtuseks saime 0.9227, millele vastav olulisustõenäosus $p\text{-}value = 0.4151$. Seega oleme kohustatud jääma normaaljaotuse eelduse juurde.

Testime Wilk-Šapiro testi töökindlust. Selleks tekitame juhuslike arvude generaatoriga n-elemendilised valimid, mis alluvad eksponentjaotusele $\mathcal{E}(1)$. Testime neid valimeid Wilk-Šapiro testiga. Antud juhul peab see test normaaljaotuse eelduse ümber lükkama. Teeme eksperimendi, et uurida valimi mahtu alates millest toimib see kummutamine praktiliselt. Teststatistikule vastavad olulisustõenäosused on toodud järgnevas tabelis:

Valimi maht n	10	15	20
$p\text{-}value$	0.59	0.021	0.0022

Tabelist võib järeldada, et Wilk-Šapiro test hakkab andma adekvaatset infot, kui valimi maht $n \geq 15$.

Järgnevalt võtame kokku kõik ülalkirjeldatud mitteparameetrilised testid. Näitame, millist statistilise analüüsi tüüpi mingi test esindab ning

milline on selle testi analoog klassikalises statistikas. See info on toodud järgmises tabelis.

Tabel 3.5. Erinevate mitteparameetriliste testide rakendamine

Analüüsi tüüp	**Mitteparameetriline test**	**Klassikalise statistika analoog**
Kahe sõltuva valimi võrdlemine	Wilcoxoni test	Kordusmõõtmiste t-test
Kahe sõltumatu valimi võrdlemine	Kolmogorov-Smirnovi test võrdlemaks kahte empiirilist jaotust	Studenti t-test kahe erineva valimi puhul
Kolme või enama sõltuva valimi võrdlemine	Friedmanni test	Kordusmõõtmistega ANOVA
Kolme või enama sõltuva valimi võrdlemine	Kruskal-Wallise H-test	Ühefaktoriline ANOVA
Kahe tunnuse vahelise seose uurimine	Spearmanni ρ või Kendalli τ	Üldine lineaarne mudel

3.5. Taasvaliku meetodid

Selles osas käsitleme statistilise anaüüsi meetodeid, mis põhinevad valimite ümberjärjestamisel. Nendest kaks kõige tuntumat on *jackknife*- ja *bootstrap*-meetodid. Mõlema meetodi idee seisneb taasvalikute tegemisel hindamaks meid huvitavat parameetrit θ (keskväärtus, mediaan, standardhälve jms). Uurime, mida see taasvalik endast täpsemalt kujutab.

3.5.1. *Jackknife*-meetod

Praktikas puutume sageli kokku heterogeensete valimitega. Säärase valimi puhul ei saa me eeldada, et kõik selle valmi elemendid alluvad samale jaotusele. Kirjeldame mõningaid selliseid olukordi.

1) Mingi sotsiaal-poliitilise suuruse (näiteks sissetulek, erakonna eelistus jms) kohta tehakse küsitlus erinevates Eesti maakondades. On ilmne, et nende näitajate jaotused erinevad piirkonniti.

2) Mõõdetakse mingi keemilise reaktsiooni kiiruskonstanti erinevatel meetoditel. On aga teada, et mõõtmisvead erinevad meetodite lõikes.

3) Tahetakse uurida hoonete soojustuse näitajaid. Kuid nende jaotused on erinevad palkmajades ning kivimajades.

Meetod nimega *jacknife* tekkis vajadusest kirjeldada andmetes peituvat heterogeensust ülaltoodud olukordades. Meetodi eestikeelne nimetus tähendab liigendnuga. Me tükeldame selle „noaga“ heterogeense valimi m homogeenseks (ehk samale jaotusele alluvaks) alamvalimiks.

Meetodi *jacknife* algoritm on järgmine.

1) Leiame valimi $\mathbf{X} = (X_1, X_2, ..., X_n)$ põhjal meid huvitava karakteristiku (keskväärtus, standardhälve jms) hinnangu $\widehat{\theta}$.

2) Jagame valimi m rühmaks ehk m homogeenseks alamvalimiks. Leiame hinnangud $\widehat{\theta}_{(j)}$, $j = 1, 2, ..., m$, kus iga kord on välja jäetud j-ndas rühm.

3) Toome sisse ja arvutame pseudoväärtused

$$\theta_{*j} = m\widehat{\theta} - (m-1)\widehat{\theta}_{(j)},$$

$j = 1, 2, ..., m$.

4) Leiame hinnangu suurusele θ :

$$\theta^* = \frac{1}{m}\sum_{j=1}^{m}\theta_{*j}$$

ning selle dispersiooni hinnangu $s_*^2 = \dfrac{s^2}{m}$, kus

$$s^2 = \frac{1}{m-1}\left(\sum_{j=1}^{m}\theta_{*j}^2 - \Big(\sum_{j=1}^{m}\theta_{*j}\Big)^2\right).$$

5) Vaatame leitud arve θ_{*j} kui normaaljaotusega juhusliku suuruse väärtusi ning leime t-jaotuse abil α-usaldusintervalli parameetrile θ :

$$\overline{\theta} = \theta^* + t_{m-1,\frac{1+\alpha}{2}}\, s_*;$$

$$\underline{\theta} = \theta^* - t_{m-1,\frac{1+\alpha}{2}}\, s_*,$$

kus suurus $t_{m-1,\frac{1+\alpha}{2}}$ on t-jaotuse $\dfrac{1+\alpha}{2}$-kvantiil vabadusastmete arvu $m-1$ korral.

Hinnangu nihke leidmine *jacknife*-meetodil

Meetodi *jacknife* abil saab leida hinnangu nihet, mida põhjustab statistik $\widehat{\theta} = T(\mathbf{X})$. Selleks moodustatakse valimi $\mathbf{X}$ baasil valimid

$$\mathbf{X}_{(i)} = (X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n)^\intercal, \quad i = 1, 2, ..., n,$$

millest on eemaldatud i-ndas vaatlus. Nende valimite põhjal leitakse hinnangud $\widehat{\theta}_{(i)} = T(\mathbf{X}_{(i)})$. Meetodile *jacknife* vastav hinnangu nihe defineeritakse kui

$$\widehat{b}^{jack} = (n-1)(\widehat{\theta}_{(.)} - \widehat{\theta}),$$

kus

$$\widehat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}.$$

Näide 3.14. Olgu meil 8-elemendilised sõltumatud valimid

$$\mathbf{X} = (2.4, 2.1, 3.5, 1.8, 1.2, 1.9, 2.1, 3)^\intercal$$

ning

$$\mathbf{Y} = (2.4, 2.8, 2.5, 1.7, 1.2, 2.1, 2.3, 2.9)^\intercal.$$

Nendest valimitest moodustatakse valim $\mathbf{Z}$, mille element $z_i = \dfrac{x_i}{y_i}$. Meid huvitab valimi $\mathbf{Z}$ põhjal saadud keskmise hinnang. Selleks moodustatakse statistik

$$\widehat{\theta} = \frac{\overline{x}}{\overline{y}} = 1.005587.$$

Kuid see hinnang on nihkega. Tõepoolest, rakendades Jenseni võrratust, saame, et

$$E(\widehat{\theta}) = E\left(\frac{\overline{x}}{\overline{y}}\right) = E(\overline{x})E\left(\frac{1}{\overline{y}}\right) \geq E(\overline{x})\frac{1}{E(\overline{y})} = \frac{E(X)}{E(Y)}.$$

Olgu selgituseks öeldud, et Jenseni võrratuse kohaselt

$$E\left(\frac{1}{\overline{y}}\right) \geq \frac{1}{E(\overline{y})}.$$

Seda võrratust saab üldistada suvalisele kumerale funktsioonile φ

$$E(\varphi(X)) \geq \varphi(E(X)).$$

Sageli on hinnangu nihe tingitud Jenseni võrratusest. Leidmaks antud hinnangu nihet arvutatakse valimite $\mathbf{Z}_{(i)}$ põhjal aritmeetilised keskmised $\widehat{\theta}_{(i)}$, $i = 1, 2, ..., 8$. Valim $\mathbf{Z}_{(i)}$ saadakse i-nda elemendi eemaldamisel valimist $\mathbf{Z}$. Saadi järgmised tulemused:

$\widehat{\theta}_{(1)}$	1.0087302
$\widehat{\theta}_{(2)}$	1.0444445
$\widehat{\theta}_{(3)}$	0.9515874
$\widehat{\theta}_{(4)}$	1.0003269
$\widehat{\theta}_{(5)}$	1.0087302
$\widehat{\theta}_{(6)}$	1.0223357
$\widehat{\theta}_{(7)}$	1.0211526
$\widehat{\theta}_{(8)}$	1.0038041

Nende põhjal saadakse, et $\widehat{\theta}_{(.)} = 1.007639$. Seega hinnangu $\widehat{\theta}$ nihe

$$\widehat{b}^{jack} = 7 \cdot (1.007639 - 1.005587) \approx 0.0144.$$

Märkus. Meetod *jacknife* ei sobi mediaani ega muude kvantiilide hindamiseks.

3.5.2. *Bootstrap*-meetod

Tänapäeval on *bootstrap*-meetod erinevates teadusharudes palju kasutusel. Meetod on pärit 1970. aastatest. Esimest korda võttis termini *bootstrap* kasutusele Stanfordi Ülikooli professor Bradley Efron, kes kirjeldas seda meetodit põhjalikult artiklis [8]. Veelgi põhjalikum ülevaade nüüdisajal rakendust leidvatest *bootstrap*-meetoditest on toodud monograafias [9]. Meetodi nimi tuleb ingliskeelsetest sõnadest *boot* (saabas) ning *starp* (paelad). Meetodi *bootstrap* ajaloolise tõlgenduse esimene versioon pärineb parun Münchauseni lugudest, kus peakangelane üritas end saapapaelu pidi soost välja tõmmata.

Matemaatilise statistika keeles tähendab see seda, et meil on lootusetuna näivad andmed. Ehk andmed, mille kohta ei oska me algselt mitte mingeid eeldusi teha. Kuid me püüame neid eeldusi tekitada ehk statistikut

andmete mädasoost välja aidata. Teine versioon on seotud Inglise armee omaaegsete saabastega, millel oli kanna-aas, kust oli lihtne sõrme läbi pistes saabas jalga tõmmata. Sellest tekkis inglise keeles kõnekäänd „*help oneself by bootstrap*“, mis sai tähenduse „*self-made man*“. See tähendab, et statistik saab oma probleemidega ise edukalt hakkama *bootstrap*'i abil.

Esmalt pisut *bootstrap*-meetodi põhimõttest. See seisneb selles, et toorvalimist $\mathbf{X} = (X_1, X_2, ..., X_n)$ moodustatakse uus valim $\mathbf{X}^*$, mille kohta saab eeldata, et ta on mingi jaotusega $\widehat{F}$. Seda uut valimit nimetatakse *bootstrap*-valimiks. Selle valimi põhjal moodustatakse statistik $T(\mathbf{X}^*, \widehat{F})$ hindamaks toorvalimit iseloomustavat parameetrit θ. Järgnevalt kirjeldame erinevaid meetodeid leidmaks *bootstrap*-valimit ning tema jaotust $\widehat{F}$.

Empiiriliselt korduv simuleerimine

Tegemist on kõige laialdasemalt levinud *bootstrap*-meetodiga. Meie sihiks on leida valimi $\mathbf{X}$ põhjal α-usaldusintervall meid huvitavale parameetrile θ.

Moodustame sellest valimist *bootstrap*-valimi $\mathbf{X}_1^*$ mahuga n järgmiselt:

1) valik on tagasipanekuga;

2) igal elemendil on valimisse kaasamise tõenäosus $\dfrac{1}{n}$.

Seega on rahuldatud valimisse kaasamise puhul klassikalise statistika eeldus 2°. Saadud valimit $\mathbf{X}_1^*$ nimetatakse *bootstrap*-valimiks. Selle valimi põhjal leitakse statistik

$$\widehat{\theta}_1^* = T(\mathbf{X}_1^*, \widehat{F}),$$

kus $\widehat{F}$ on antud juhul diskreetne ühtlane jaotus. Kordame kirjeldatud valimi võtmist ja statistiku leidmist m korda. Saame *bootstrap*-valimid $X_1^*, X_2*, ..., X_m^*$ ning nende põhjal leitud statistikud $\widehat{\theta}_1^*, \widehat{\theta}_2^*, ..., \widehat{\theta}_m^*$. Seejärel leitakse nende statistikute moodustatud empiiriline jaotus. Selle jaotuse põhjal saame leida hinnangud suuruse X keskväärtusele, standardhälbele, asümmeetriale ning meid huvitavatele kvantiilidele. Näiteks

standardhälbe hinnang

$$\widehat{s} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\theta}_i^* - \overline{\theta^*})^2},$$

kus

$$\overline{\theta^*} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\theta}_i^*.$$

Samuti saame *bootstrap*-meetodiga hinnata valimi keskväärtuse $E(X) = \mu$ hinnangu nihet. Algse valimi põhjal saame nihke hinnangu

$$\widehat{b} = \overline{X} - \mu = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu.$$

Antud juhul aga pole teada infot valimi jaotuse kohta, seega ei saa me hinnata keskväärtust μ otsesel kujul. Vaatame keskväärtuse hinnangut *bootstrap*-valimite korral. Valimi X_i^* põhjal leitud aritmeetiline keskmine $\overline{x}^*$ on õige keskväärtus, seega $E(X_i*) = \overline{x}^*$, $i = 1, 2, ..., m$. Leides *bootstrap*-hinnangud $\overline{x_i}^*$, saame konstrueerida nihke hinnangu

$$\widehat{b}^* = \frac{1}{m}\sum_{i=1}^{m}\overline{x_i}^* - \overline{x} = \widehat{\overline{X} - \mu}.$$

Märkus. Empiiriliselt korduv simuleerimine sobib valimi keskmise väärtuse hindamiseks. Miinimumide ja maksimumide hindamiseks tuleb rakendada muid meetodeid.

Parameetriline *bootstrap*

Parameetriline *bootstrap* on olemuselt küllaltki sarnane suurima tõepära (STP) meetodile. Nimelt on parameetrilise *bootstrap*-meetodi eesmärk kaasata võimalikult palju valimis peituvat infot hindamaks meid huvitavat karakteristikut. Meetodi rakendamise etapid on järgmised.

1) Hindame algse valimi **X** põhjal erinevaid jaotusi. Valime neist välja sellise, mis sobib antud valimiga kõige paremini.

2) Leiame valitud jaotuse parameetritele STP hinnangud.

3) Tekitame juhuslike arvude generaatoriga uued valimid valitud jaotuse ning hinnatud parameetrite baasil.

4) Leiame nende uute valimite põhjal parameetrile θ *bootstrap*-hinnangu.

Meetodi eelis on, et ta ei lähtu hinnangu tegemisel mitte üksnes keskmistest väärtustest, vaid ka väikestest ning suurtest väärtustest. Kuid meetod eeldab küllaltki täpset eelinfot uuritava suuruse jaotusest. Seega tekib küsimus, kuidas leida algse valimi põhjal sobivaim jaotus. Ühe võimaliku vastuse sellele küsimusele andis Efron artiklis [10]. Selles töös soovitatakse lähtuda kas mitmemõõtmelisest normaaljaotusest või seosega
(2.10) defineeritud jaotuste perest (eksponentsiaalsete jaotuste perest).

3.5.3. Vahemikhinnangute leidmine *bootstrap*-meetodil

Klassikalisel juhul moodustasime huvipakkuvale parameetrile θ sümmeetrilise α-usaldusintervalli. Selleks leidsime talle statistiku $T(X)$ abil hinnangu $\widehat{\theta}$ ning selle hinnangu standardhälbe $\sigma_{\widehat{\theta}}$. Saadud α-usaldusintervall

$$I_\alpha = [\widehat{\theta} - \sigma_{\widehat{\theta}} z_\alpha; \widehat{\theta} + \sigma_{\widehat{\theta}} z_\alpha],$$

kus z_α tähistab kriitilist väärtust (ehk $\dfrac{1+\alpha}{2}$ -kvantiile) kasutatava jaotuse (peamiselt normaaljaotuse või t-jaotuse) korral.

Säärane lähenemine andis adekvaatse tulemuse kahel juhul.

1) Uuritav suurus allus normaaljaotusele.

2) Toimis tsentraalne piirteoreem.

Kui aga on teada, et meie valimi jaotus on ebasümmeetriline, siis annab klassikaline lähenemine mitteadekvaatse vahemikhinnangu. Sel juhul on võimalik *bootstrap*-meetodi abil muuta α-usaldusintervalli realistlikumaks.

Empiirilisel jaotusfunktsioonil põhinev vahemikhinnang

Tegemist on kõige loomulikuma meetodiga leidmaks α-usaldusintervalli *bootstrap*-meetodil. Olgu meil m *bootstrap*-valimit $\mathbf{X}_i^*$ ning igaühe puhul leitud parameetri θ hinnang $\widehat{\theta}_i^*$. Olgu $F_m(x)$ suuruste $\widehat{\theta}_i^*$ empiiriline jaotusfunktsioon. Siis avaldub sümmeetrilise jaotuse puhul α-usaldusintervall parameetrile θ kujul

$$I_\alpha = \left[\underline{\theta}_\alpha; \overline{\theta}_\alpha\right] = \left[F_m^{-1}\left(\frac{1-\alpha}{2}\right); F_m^{-1}\left(\frac{1+\alpha}{2}\right)\right].$$

Seega peitub meetodi rakendamise kogu raskus $\frac{1-\alpha}{2}$- ja $\frac{1+\alpha}{2}$-kvantiilide leidmises juhuslikule suurusele $\widehat{\theta}^*$. Meetodi oluliseks plussiks on, et usaldusvahemik jääb alati lubatud muutumispiirkonda. Näiteks osakaalu usaldusvahemiku alumine raja on alati positiivne ja ülemine raja ei ole suurem kui 1.

Bootstrap'i t-meetod ehk studentiseerimine

See meetod järgib teatud mõttes klassikalist lähenemist usaldusintervallide leidmisel. Lihtsamal erijuhul on tegemist tsentraalse piirteoreemi rakendamisega.

Olgu meil m *bootstrap*-valimit. Leiame iga valimi korral statistiku

$$Z_i^* = \frac{\widehat{\theta}_i^* - \widehat{\theta}}{\widehat{s}_{\theta_i}^*}, \qquad i = 1, 2, ..., m,$$

kus $\widehat{\theta}_i^* = T(\mathbf{X}_i^*)$ on i-nda *bootstrap*-valimi $\mathbf{X}_i^*$ põhjal leitud hinnang meid huvitavale suurusele θ ning $\widehat{s}_{\theta_i}^*$ selle hinnangu standardhälbe hinnang. Selle standardhälbe hinnangu leidmiseks tuleb täiendavalt genereerida m_i *bootstrap*-valimit valimist $\mathbf{X}_i^*$. Enamasti piisab, et $m_i = 50$. Järgmise sammuna leiame juhusliku suuruse Z_i^* jaoks α-kvantiili $\widehat{t}_\alpha$ tingimusest

$$\frac{\mid \{z_i^* \mid z_i^* \leq \widehat{t}_\alpha\} \mid}{m} = \alpha.$$

Siis saame järgmise tingimuse leidmaks α usaldusintervalli suurusele θ :

$$P\left(\widehat{t}_{\frac{1-\alpha}{2}} \leq \frac{T(\mathbf{X}) - \theta}{s_\theta} \leq \widehat{t}_{\frac{1+\alpha}{2}}\right) \approx \alpha,$$

kus statistik $T(\mathbf{X})$ on algvalimi põhjal saadud hinnang suurusele θ. Selle hinnangu standardhälbe hinnang s_θ leitakse *bootstrap*-hinnangute põhjal järgmiselt:

$$s_\theta = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\theta_i^* - \frac{1}{m}\sum_{j=1}^{m}\theta_j^*\right)^2}.$$

Seega saame meid huvitava usaldusintervalli

$$I_\alpha \approx [\widehat{\theta} - s_\theta \widehat{t}_{\frac{1+\alpha}{2}}; \widehat{\theta} - s_\theta \widehat{t}_{\frac{1-\alpha}{2}}].$$

Kui algvalimi maht $n \geq 100$ ning selle valimi iga element allub samale jaotusele, siis saame studentiseerimise meetodi taandada tsentraalse piirteoreemi rakendamisele. Selleks leiame empiiriliselt korduva simuleerimise teel m *bootstrap*-valimit ning iga selle valimi põhjal hinnangud

$$\overline{x_i^*} = \sum_{j=1}^{n} X_j^*, \qquad i = 1, 2, ..., m.$$

Kui hinnatavaks suuruseks θ on algvalimi keskväärtus, siis statistik

$$Z_i^* = \frac{\overline{x_i^*} - \theta}{s_\theta}$$

allub ligikaudselt standardsele normaaljaotusele. Saame järgmise 0.95-usaldusintervalli parameetrile θ :

$$I_{0.95} \approx [\widehat{\theta} - 1.96 s_\theta; \widehat{\theta} + 1.96 s_\theta].$$

Bootstrap'i t-meetodil on 2 olulist puudust.

1) Usaldusvahemik võib olla muutumispiirkonnast väljas. Võib näiteks juhtuda, et osakaalu hinnangu alumine piir on negatiivne, ülemine aga suurem kui 1.

2) Meetod on väga töömahukas. Kokku tuleb meil genereerida $m \cdot m_i$ *bootstrap*-valimit.

Korrigeeritud nihkega usaldusintervall

Vahemikhinnangu puhul on sageli probleemiks hinnangu nihe ning uuritava suuruse asümmeetria. Sel juhul aitab sageli meetod, kus korrigeeritakse usaldusintervalli nihke- ning asümmeetriaparanditega. Inglise keeles on selle meetodi nimi BC_a*-method* (***B**ias **C**orrected and **a**ccelerated*

method). Lühidalt võib meetodit nimetada kui $\mathrm{BC_a}$. Selle meetodi rakendamisel tuuakse sisse kaks täiendavat konstanti:

1) asümmeetriaparand $\widehat{a}$ (ingl *acceleration*),

2) nihkeparand $\widehat{b}$ (ingl *bias-correction*).

Meetod $\mathrm{BC_a}$ toimib valimitel, mille puhul on pisut rikutud normaaljaotuse või tsentraalse piirteoreemi eeldusei.

Nihke- ja asümmeetriaparandit rakendades saame järgmise eeskirja konstrueerimaks α-usaldusintervalli meid huvitavale parameetrile θ :

$$I_\alpha = [\theta_{\alpha_1}; \theta_{\alpha_2}],$$

kus

$$\alpha_1 = \Phi\left(\widehat{b} + \frac{\widehat{b} + z_{\frac{1-\alpha}{2}}}{1 - \widehat{a}(\widehat{b} + z_{\frac{1-\alpha}{2}})}\right) + \frac{1}{2}, \tag{3.10}$$

$$\alpha_2 = \Phi\left(\widehat{b} + \frac{\widehat{b} + z_{\frac{1+\alpha}{2}}}{1 - \widehat{a}(\widehat{b} + z_{\frac{1+\alpha}{2}})}\right) + \frac{1}{2}. \tag{3.11}$$

Funktsioon Φ seostes (3.10) ja (3.11) on Laplace'i veafunktsioon ning $z_{\frac{1+\alpha}{2}}$ on standardse normaaljaotuse $\dfrac{1+\alpha}{2}$ -kvantiil. Kui $\widehat{b} = 0$ ning $\widehat{a} = 0$, siis

$$\alpha_1 = \Phi(z_{\frac{1+\alpha}{2}}) + \frac{1}{2} \text{ ja } \alpha_2 = \Phi(z_{\frac{1-\alpha}{2}}) + \frac{1}{2}.$$

Vaatame, kuidas määrata $\widehat{a}$ ja $\widehat{b}$ võrdustes (3.10)-(3.11). Nihkeparand $\widehat{b}$ mõõdab parameetri θ hinnangu nihet ning saadakse võrdusest

$$\widehat{b} = \Phi^{-1}\left(\frac{\mid \{\theta_i^* \mid \theta_i^* < \widehat{\theta}\} \mid}{m} - \frac{1}{2}\right),$$

kus $\widehat{\theta}$ on esialgse valimi põhjal leitud hinnang ning θ_i^*, $i = 1, 2, ..., m$ *bootstrap*-valimi põhjal leitud hinnangud. Kui $\dfrac{m}{2}$ väärtuste puhul $\theta_i^* < \widehat{\theta}$, siis $\widehat{b} = 0$, sest $\Phi^{-1}(0) = 0$.

Asümmeetriaparandi $\widehat{a}$ leidmine on tunduvalt töömahukam. Üks võimalusi selle leidmiseks on kasutada *jacknife*-meetodi ideed. Olgu $\mathbf{X}_{(i)}$ esialgne valim, kust on eemaldatud i-ndas vaatlus ning suurus $\widehat{\theta}_{(i)}$ parameetri θ hinnang selle valimi põhjal. Olgu

$$\widehat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}.$$

Siis saame asümmeetriaparandiks

$$\widehat{a} = \frac{\sum_{i=1}^{n}(\widehat{\theta}_{(\bullet)} - \widehat{\theta}_{(i)})^3}{6\left(\sum_{i=1}^{n}(\widehat{\theta}_{(\bullet)} - \widehat{\theta}_{(i)})^2\right)^{\frac{3}{2}}}.$$

Blokk-*bootstrap*

Meetodi ingliskeelne nimetus on *block bootstrap*. Ülaltoodud meetodite rakendamisel kaotasime info vaatlustevahelistest sõltuvustest. Seega ei sobi need meetodid aegridadele, kus sõltuvus ajast on oluline. Selleks puhuks sobib blokk-*bootstrap*'i meetod.

Meetodi realiseerimiseks jagatakse vaatlused esiteks blokkidesse, mille pikkus l suureneb valimi mahu suurenedes. Rusikareegli kohaselt

$$l = \sqrt[3]{n}.$$

Antud juhul võib suurust n käsitleda kui aegrea pikkust. Valimi blokid võivad olla nii ühisosata kui ka ühisosaga nagu näiteks

$$\{X_1, X_2, ..., X_l\}, \{X_2, X_3, ..., X_{l+1}\}, ..., \{X_{n-l+1}, X_{n-l+2}, ..., X_n\}.$$

Seejärel rakendatakse igale blokile taasvaliku meetodit ning moodustatakse *bootstrap*-valimid. Nende valimite põhjal leitakse regressioonikordajate hinnangud.

Blokk-*bootstrap*'i kohta võib huviline põhjalikumalt teada saada näiteks monograafiatest [24] või [7].

3.5.4. *Bootstrap*-hinnangud tarkvara R abil

Toome näiteid tarkvara R mõningatest programmidest, mille abil saab leida *bootstrap*-hinnanguid. Esmalt vaatame enim kasutatud *bootstrap*-meetodit, milleks on empiiriliselt korduv simuleerimine. Üks võimalusi on seda simuleerimist teha allpool järgneva programmiga:

```
x=c(4,7,15.1,12.3,6)
y=0
for (i in 1:1000)
y[i]=mean(sample(x,length(x),replace=TRUE,prob=NULL))
u=quantile(y,0.025)
U=quantile(y,0.975).
```

Antud juhul on meil valimi $\mathbf{X}$ realisatsioon $\mathbf{x} = (4, 7, 15.1, 12.3, 6)$. Sellest valimist moodustati 5-elemendiline *bootstrap*-valim $\mathbf{Y}$ selliselt, et valik oleks tagasipanekuga ning igal elemendil oleks valimisse kaasamise tõenäosus $\frac{1}{5}$. Kokku genereeriti 1000 sellist valimit ning iga valimi põhjal leiti aritmeetilised keskmised $\overline{y}_i$, $i = 1, 2, ..., 1000$. Lõpuks leiti nende aritmeetiliste keskmiste empiirilise jaotuse 0.025-kvantiil u ja 0.975-kvantiil U. Tulemuseks saadi uuritava suuruse keskväärtuse 0.95-usaldusintervalli hinnang $[u, U]$.

Teiseks uurime meetodi $\mathrm{BC_a}$ realiseerimist tarkvara R abil. Teeme seda 3 etapis.

Esimene etapp. Esimeseks etapiks on valimi realisatsiooni (mõõtmistulemuste) sisestamine ning hinnangu $\widehat{\theta}$ leidmine. Seda teeb järgmine programm:

```
x=cbind(8.1,7.6,10.1,12.4,7.8,13.1,7.9,15.5,8.2)
keskm=mean(x).
```

Hinnanguks $\widehat{\theta}$ on selles programmiosas tunnus `keskm`.

Teine etapp. Teises etapis leitakse asümmeetriaparand $\widehat{a}$. Selle parandi leiab järgmine programmiosa:

```
a=0
for (i in 1:length(x)) a[i]=mean(x[-i])
kesk=mean(a)
```

```
k1=0
for (i in 1:length(a)) k1=k1+(kesk-a[i])^3
k2=0
for (i in 1:length(a)) k2=k1+(kesk-a[i])^2
k2=6*k2^(1.5)
asym=k1/k2-
```

Tunnus `a[i]` selles programmiosas tähistab hinnanguid $\widehat{\theta}_{(i)}$, $i = 1, 2, ..., n$. Hinnangu $\widehat{\theta}_{(\bullet)}$ jaoks on tunnus `kesk`. Asümmeetriaparandiks $\widehat{a}$ aga on tunnus `asym`.

Kolmas etapp. Viimases etapis leitakse nihkeparand $\widehat{b}$ ning korrigeeritud α-usaldusintervall I_α. Selleks on järgmine programm:

```
b=0
for (j in 1:1000)
b[j]=mean(sample(x,length(x),replace=TRUE,prob=NULL))
k=length(sample(b[b<=keskm]))
nihe=qnorm(k/length(b))
alpha1=pnorm(nihe+(nihe+qnorm(0.025))/
(1-asym*(nihe+qnorm(0.025))))
alpha2=pnorm(nihe+(nihe+qnorm(0.975))/
(1-asym*(nihe+qnorm(0.975))))
alumuus=quantile(x,alpha1)
ylemuus=quantile(x,alpha2).
```

Nihkeparameetriks $\widehat{b}$ on selles programmiosas tunnus `nihe`. Tunnused `alumuus` ning `ylemuus` tähistavad parameetri θ korrigeeritud nihkega α-usaldusintervalli piire u_{α_1} ning U_{α_2}. Antud juhul on tegemist 0.95-usaldusintervalliga.

Ülaltoodud programm töötab hästi juhul, kui uuritava suuruse jaotus kaldub mõõdukalt kõrvale normaaljaotusest. Muude jaotuste juhtudel kipub ta andma ebaadekvaatseid tulemusi. Üldisema $\mathrm{BC_a}$ meetodi jaoks saab vastava R-i paketi installeerida alljärgneva käsuga:

```
install.packages("bootBCa",
repos="http://R-Forge.R-project.org").
```

Kui see pakett on alla laaditud, siis saab leida 0.95-usaldusintervalli keskväärtusele järgmiselt:

`BCa(x,NA,mean,alpha=c(0.025,0.975),M=1000)`.

Algandmete vektorit tähistab `x` ning argument `M=1000` määrab *bootstrap*-valimite hulgaks 1000. Käsuga `BCa` saab hinnata ka dispersiooni (`mean` asemel `var`) ja mediaani (`mean` asemel `median`) usaldusintervalle.

3.5.5. Permutatsiooni test

Permutatsiooni testi näol on tegemist sellise statistilise testiga, mille puhul saadakse teststatistiku väärtused andmete erinevate ümberjärjestuse korral. Lahti seletatult: olemasolevatest valimitest moodustatakse taasvaliku meetodil uued valimid ning leitakse nende põhjal meid huvitavale statistikule väärtused. Seda valikut korratakse m korda. Iga kord eeldatakse statistiku väärtuse leidmisel nullhüpoteesi kehtimist.

Demonstreerime permutatsiooni testi ideed 2 valimi abil. Olgu meil valimid

$$\mathbf{X} = (X_1, X_2, ..., X_k)^\intercal \text{ ning } \mathbf{Y} = (y_1, y_2, ..., y_l)^\intercal,$$

$k+l = n$. Eeldame, et need valimid on sõltumatud ning vastavalt jaotustega F ja G. Meie eesmärk on valimite realisatsioonide $\mathbf{x}$ ning $\mathbf{y}$ põhjal testida nullhüpoteesi, mille korral ei ole erinevust jaotuste F ja G vahel. Seega

$$H_0 : F = G.$$

Antud juhul tähendab võrdusmärk, et iga x_i ning y_j, $i = 1, 2, ..., k$ ning $j = 1, 2, ..., l$ korral

$$P(X \leq x_i) = P(Y \leq y_j).$$

Olgu meil järgmised valimite $\mathbf{X}$ ning $\mathbf{Y}$ realisatsioonid:

$$\mathbf{x} = (94, 197, 16, 38, 99, 141, 23)^\intercal,$$

$$\mathbf{y} = (52, 104, 146, 10, 50, 31, 40, 27, 46)^\intercal.$$

Antud juhul $k = 7$ ja $l = 8$. Permutatsiooni testi läbiviimiseks moodustatakse järkstatistikute väärtustest 16-komponendiline vektor $\mathbf{z} = (\mathbf{x}, \mathbf{y})^\intercal$. Vektorit $\mathbf{z}$ kirjeldab järgmine tabel:

Valim	x	y	y	x	x	y	x	x	x	x	y	y	x	y	x	y
Astak	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Väärtus	10	16	23	27	31	38	40	46	50	52	94	99	104	141	146	197

Kõik 16 vaatlust on järjestatud vähimast suurimani. Iga vaatlus kuulub kas valimi realisatsiooni $\mathbf{x}$ või $\mathbf{y}$. Olgu vektor $\mathbf{v} = (10, 16, 23, ..., 197)^\intercal$ ning vektor $\mathbf{g} = (g_1, g_2, ..., g_n)^\intercal$, mille komponent

$$g_i = \begin{cases} x, & \text{kui vaatlus kuulub valimisse } \mathbf{X}, \\ y, & \text{kui vaatlus kuulub valimisse } \mathbf{Y}. \end{cases}$$

Seega sisaldavad vektorid $\mathbf{v}$ ning $\mathbf{g}$ kogu informatsiooni vektori $\mathbf{z} = (\mathbf{x}, \mathbf{y})^\intercal$ kohta. Vektori $\mathbf{g}$ komponentide hulk on n, millest k on väärtusega $\mathbf{x}$ ning l väärtusega $\mathbf{y}$. Kokku on

$$C_n^k = \frac{n!}{k!l!}$$

võimalust jagada vektori $\mathbf{g}$ komponendid alamhulkadesse väärtustega $\mathbf{x}$ ning $\mathbf{y}$. Permutatsiooni test põhineb järgmisel tulemusel.

Permutatsiooni lemma. Nullhüpoteesi $H_0 : F = G$ korral on vektori $\mathbf{g}$ iga alamhulkade $\mathbf{x}$ ja $\mathbf{y}$ kombinatsiooni esinemise tõenäosus $\dfrac{1}{C_n^k}$.

Moodustame teststatistiku $\widehat{\theta}$ kui vektorite $\mathbf{g}$ ning $\mathbf{v}$ funktsiooni

$$\widehat{\theta} = T(\mathbf{g}, \mathbf{v}).$$

Uurime lähemalt statistikut $\widehat{\theta} = \overline{x} - \overline{y}$, mida saab esitada vektori $\mathbf{g}$ kaudu kujul

$$\widehat{\theta} = \frac{1}{k} \sum_{g_i = x} v_i - \frac{1}{l} \sum_{g_i = y} v_i.$$

Selle teststatistiku korral on tegemist permutatsiooni testiga, mis üldistab Studenti t-testi. Statistik $\widehat{\theta}$ esindab ülaltoodud tabelis esitatud algseisu. Olgu $\mathbf{g}^*$ üks võimalikest alamhulkade x ja y kombinatsioonidest moodustatud vektor. Nende vektorite põhjal saab leida statistikud

$$\widehat{\theta}^* = \widehat{\theta}(\mathbf{g}^*) = T(\mathbf{g}^*, \mathbf{v}).$$

Kõikvõimalikke vektoreid $\mathbf{g}^*$ ning statistikuid $\widehat{\theta}^*$ on kokku C_n^k. Kõikidest leitud statistikutest $\widehat{\theta}^*$ moodustub meid huvitava teststatistiku jaotus nullhüpoteesi H_0 korral. Selle jaotuse põhjal leitakse tõenäosus nimega

ASL (ingl ***A**chieved **S**ignificance **L**evel*). Eesti keeles tähendab see saavutatud olulisuse nivood. Antud juhul

$$\text{ASL} = P(\widehat{\theta}^* > \widehat{\theta}) = \frac{\mid \{\widehat{\theta}^* > \widehat{\theta}\} \mid}{C_n^k}.$$

Tõenäosust ASL võib käsitleda kui olulisustõenäosuse *p-value* robustset hinnangut. Samuti saab suuruse ASL abil hinnata valitud statistiku töökindlust. Üldiselt

$$P(\text{ASL} \leq \alpha) = \alpha,$$

iga $\alpha \in (0; 1)$.

Permutatsiooni testi võib võtta kokku järgmise algoritmiga.

1) Moodustame m erinevat vektorit $\mathbf{g}^*(1), \mathbf{g}^*(2), ..., \mathbf{g}^*(m)$, mis moodustatakse juhuslikult erinevate valimite kõikvõimalikest kombinatsioonidest.

2) Igale kombinatsioonile leitakse vastav statistik

$$\widehat{\theta}^*(i) = T(\mathbf{g}^*(i), \mathbf{v}), \quad i = 1, 2, ..., m.$$

3) Leitakse tõenäosuse ASL hinnang

$$\widehat{\text{ASL}} = \frac{\mid \{j \mid \widehat{\theta}^*(j) > \widehat{\theta}\} \mid}{m}.$$

Permutatsiooni testi algoritm on küllaltki sarnane eespool toodud *bootstrap*-meetodi algoritmidega. Peamine erinevus seisneb selles, et valik on tagasipanekuta. See tähendab, et peale elemendi valimisse kaasamist ei saa teda enam valida. Teisisõnu tähendab tagasipanekuta valik klassikalise statistika sõltumatuse eelduse rikutust.

Tekib küsimus, kui suur peaks olema korduste hulk m. Vastamaks sellele küsimusele toome sisse ühe uue suuruse. Olgu $A = \text{ASL}$ ning $\widehat{A} = \widehat{ASL}$. Siis võrdub $m\widehat{A}$ statistikute $\widehat{\theta}^*$ hulgaga, mille hinnangulised väärtused ületavad algvalimi põhjal leitud statistiku $\widehat{\theta}$ väärtust. Antud juhul

$$m\widehat{A} \sim B(m, A); \quad E(\widehat{A}) = A \text{ ning } D(\widehat{A}) = \frac{A(1-A)}{m}.$$

Defineerime suuruse

$$\mathrm{cv}(\widehat{A}) = \sqrt{\frac{1-A}{mA}},$$

mida nimetatakse variatsiooni koefitsiendiks (ingl *coefficient of variance*). Korduste hulk m sõltub ette antud variatsiooni koefitsiendi suurusest. Alljärgnevalt on toodud minimaalsed vajalikud m väärtused sõltuvalt tõenäosustest ASL, kui $\mathrm{cv}(\widehat{A}) \leq 0.1$:

ASL	0.5	0.25	0.1	0.05	0.025
m	100	299	900	1901	3894

Teostamaks permutatsiooni testi tarkvaras R tuleb sellesse installeerida käsuga

```
install.packages("perm")
```

pakett `perm`. Siis saab permutatsiooni testi läbi viia järgmiselt:

```
x=c(94,197,16,38,99,141,23)
y=c(52,104,146,10,50,31,40,27,46)
DV=c(x,y)
IV <- factor(rep(c("x", "y"), c(length(x), length(y))))
permTS(DV~IV, alternative="two.sided", exact=TRUE).
```

Käsu `permTS` väljundiks on olulisustõenäosus *p-value*, mis on antud andmete korral 0.281.

Permutatsiooni testiga saab kontrollida ka valimite dispersioonide võrdsuse hüpoteesi ehk üldistada Fisheri F-testi. Selleks tuleb koostada teststatistik

$$\widehat{\theta} = \ln\left(\frac{s_x^2}{s_y^2}\right).$$

Kahepoolse hüpoteesi korral saame pärast m permutatsiooni tõenäosuse ASL hinnangu

$$\widehat{\mathrm{ASL}} = \frac{|\,\{j \mid\mid \widehat{\theta}^*(j) \mid > \mid \widehat{\theta} \mid\}\,|}{m}.$$

3.6. Ülesanded

Ülesanne 3.1. Allugu lambipirni eluiga T eksponentjaotusele. Olgu lambipirni keskmine eluiga $E(T) = 10$ kuud. Leidke 8 sõltumatult põleva lambipirni maksimaalse ja minimaalse eluea tihedusfunktsioonid.

Ülesanne 3.2. Olgu juhuslikud suurused X_1, X_2 ja X_3 sõltumatud ja ühtlase jaotusega lõigul $[0;2]$. Olgu $X_{(1)}$, $X_{(2)}$ ja $X_{(3)}$ vastavad järkstatistikud. Leidke tõenäosused $P(X_{(1)} > 0.1)$ ning $P(X_{(3)} \leq 0.9)$.

Ülesanne 3.3. Olgu meil 6-elemendiline valim sõltumatutest juhuslikest suurustest, mis on tihedusfunktsiooniga

$$f(x) = \begin{cases} 0, \text{ kui } x \notin \left[-\frac{\pi}{2}; \frac{\pi}{2}\right], \\ \frac{\cos(x)}{2}, \text{ kui } x \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right]. \end{cases}$$

Leidke järkstatistikute $X_{(1)}$ ja $X_{(6)}$ jaotusfunktsioonid $F_{X_{(1)}}$ ja $F_{X_{(6)}}$ ning tihedusfunktsioonid $f_{X_{(1)}}$ ja $f_{X_{(6)}}$.

Ülesanne 3.4. Olgu meil 9-elemendiline valim $\mathbf{X}$. Leidke järkstatistiku $X_{(5)}$ (ehk mediaani) jaotus- ja tihedusfunktsioon, kui seda valimit esindava juhusliku suuruse tihedusfunktsioon on järgmine:

$$f(x) = \begin{cases} 0, \text{ kui } x \notin [0;2], \\ \frac{x}{2}, \text{ kui } x \in [0;2]. \end{cases}$$

Ülesanne 3.5. Olgu meil juhuslikud suurused $X_1, X_2, ..., X_n$ sõltumatud ning jaotusfunktsiooniga

$$F(x) = \begin{cases} 0, \text{ kui } x < 1, \\ 1 - \left(\frac{1}{x}\right)^2, \text{ kui } x \geq 1. \end{cases}$$

Olgu $Y_n = \max\{X_1, X_2, ..., X_n\}$, $n \geq 1$. Leidke juhusliku suuruse Y_n jaotusfunktsioon, kui $n \to \infty$. Näpunäide: võtke definitsiooni 3.1 jadaks $a_n = n^{\frac{1}{2}}$ ning jadaks $b_n \equiv 0$.

Ülesanne 3.6. Olgu juhuslikud suurused X_1, X_2 ja X_3 sõltumatud ja ekponentjaotusega parameetriga 1. Leidke keskväärtus $E(X_{(2)})$. Võrrelge saadud tulemust eksponentjaotusele parameetriga 1 alluva juhusliku suuruse keskväärtuse ja mediaaniga.

Ülesanne 3.7. Olgu juhuslikud suurused $X_1, X_2, ..., X_{10}$ ühtlase jaotusega lõigul $[0; 10]$. Leidke järgmine tõenäosus:

$$P(X_{(1)} > 2, X_{(10)} \leq 8).$$

Ülesanne 3.8. Maksku kindlustusfirma tormikahjustuse summa X (tuhandetes eurodes), mida kirjeldab tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x \notin [0; 40], \\ \dfrac{3}{32000}(40x - x^2), \text{ kui } x \in [0; 40]. \end{cases}$$

Leidke 8 väljamakse põhjal saadud tõenäosus, et minimaalne kahjusumma on üle 1000 euro ja maksimaalne kahjusumma samal ajal ei ületaks 15 000 eurot.

Ülesanne 3.9. Mida järeldada Spearmanni korrelatsioonikordaja $\rho(X, Y)$ kohta järgnevatel juhtudel:

1) $y = \exp(-2x)$,

2) $y = \dfrac{x}{1+x}$,

3) $y = \ln(x^2)$, kui $x \neq 0$,

4) $y = \arccos(x)$?

Ülesanne 3.10. Olgu meil järgmine andmestik:

X	Y
1	15
2	11
3	10
4	18
8	35

Leidke tunnuste X ja Y vaheline Pearsoni lineaarne korrelatsioonikordaja, Spearmanni astakkorrelatsioonikordaja $\rho(X,Y)$ ja Kendalli $\tau(X,Y)$. Võrrelge neid kordajaid. Mida järeldate?

Ülesanne 3.11. Ligi 700 °C temperatuuri juures kuumutati mingi kindla fikseeritud aja 5 raua, boori ja räni sulami proovitükki. Pärast kuumutamist mõõdeti selle sulami tüki passiveerimise potentsiaali, s.t suurust, mis iseloomustab kristalliseerunud sulami eritakistust. Saadi järgmised tulemused:

Kuumutamise aeg, min	Passiveerimise potentsiaal, mV
10	−408
20	−400
45	−392
90	−379
120	−385

Leidke kuumutamise aja ja passiveerimise potsentsiaali vaheline Spearmanni korrelatsioonikordaja ρ ja Kendalli τ.

Ülesanne 3.12. Kuus keemikute gruppi (grupid A, B, C, D, E ja F) mõõtsid ühe esimest järku keemilise reaktsiooni kiiruskonstanti (t^{-1}). On teada, et gruppide mõõtmismetoodikad ning ka grupis olnud keemikute kvalifikatsioonid erinesid. Probleem: milline võiks olla nende andmete põhjal kiiruskonstandi hinnang ja selle hinnangu 0.95-usaldusvahemik? Võrrelge usalduspiiride *jacknife*-hinnangut klassikalise (s.t Studenti t-jaotuse abil saadud) hinnanguga. Mõõtmistulemused on toodud allpool tabelis.

A	B	C	D	E	F
1.48	2.33	1.35	2.36	2.15	1.01
1.43	0.71	1.26	0.93	2.26	1.09
1.22	1.67	1.2	1.55	1.84	1.26
1.9	1.61	1.67	1.9	2.21	1.2
1.37	1.08	1.37	1.95	1.79	1.18

Ülesanne 3.13. Valimisse võeti 8 tulekindla tsemendi proovi. Saadi järgmised jämedate graanulite protsentuaalsed sisaldumised proovides:

1.7	0.9	3.4	2.5	3.1	0.6	1.0	2.1

Kas see valim on piisav tõestamaks, et vähem kui pooltes proovides on jämedate graanulite osakaal üle 2%? Olgu testi olulisuse nivooks $\beta = 0.05$.

Ülesanne 3.14. Mõõdeti 6 detaili läbimõõtu (mm) kahel erineval moel, pooled ühel ja pooled teisel meetodil. Saadi järgmised tulemused:

2.1	1.8	1.9	2.5	2.2	1.4

Kuna meetodid olid erinevad, siis ei saa anda adekvaatset hinnangut valimi jaotusele. Leidke 500 *bootstrap*-valimi põhjal detaili keskmisele läbimõõdule 0.95-usaldusintervalli hinnang.

Ülesanne 3.15. Suhkrupeedis mõõdeti C-vitamiini sisaldust milligrammides. Saadi järgmised tulemused:

52	60	66	61	49	50	48	53

Võrrelge C-vitamiini sisalduse keskväärtusele t-jaotuse põhjal saadud 0.95-usaldusintervalli korrigeeritud nihke meetodil saadud 0.95-usaldusintervalliga. Usaldusintervalli võiks hinnata 1000 *bootstrap*-valimi põhjal.

Ülesanne 3.16. Insener sai algse 10-elemendilise valimi põhjal detaili keskmise läbimõõdu hinnanguks $\widehat{\theta} = 12$ mm. Standardhälbe abil sai insener keskmisele läbimõõdule 0.95-usaldusintervalliks $[9.95; 14.05]$. Samas aga tekkis inseneril kahtlus, et leitud hinnang on nihkega. Ta pöördus matemaatiku poole. Matemaatik tegi 200 *bootstrap*-valimit. Osutus, et 150 juhul oli *bootstrap*-hinnang väiksem kui 12 mm ning 50 juhul suurem kui 12 mm. Seda arvestades tegi matemaatik nihkega korrigeeritud vahemikhinnangu. Milliseks muutus pärast seda detaili keskmise läbimõõdu 0.95-usaldusintervall?

Ülesanne 3.17. Neli eksperti (A, B, C ja D) hindasid vastvalminud hoonet kümnepallisüsteemis. Hoone juures hinnati 6 omadust. Saadi järgmised tulemused:

Omadus	Ekspert A	Ekspert B	Ekspert C	Ekspert D
1	8	9	7	8
2	6	5	9	7
3	10	7	7	8
4	9	6	8	6
5	9	10	10	7
6	9	8	6	8

Kas võib olulisuse nivool 0.05 ümber lükata väite, et neli eksperti hindasid omadusi samaväärselt? Näpunäide: kasutada Friedmanni testi.

Ülesanne 3.18. Merevee seisundit hinnati kümnepallisüsteemis 3 erineva meetodiga. Iga meetodi puhul kasutati 6 eksperdi arvamust. Saadi järgmised hinnangud (0–10):

Meetod 1	Meetod 2	Meetod 3
10	9	10
8	7	8
9	8	8
7	6	8
10	9	10
8	9	9

Kontrollige nullhüpoteesi, mille kohaselt andsid kõik 3 meetodit sama tulemuse. Olgu olulisuse nivoo $\beta = 0.05$.

Ülesanne 3.19. Üle 100 kg kaaluvate meeste peal katsetati 2 erinevat kaalu langetamise metoodikat (metoodikat A ja B). Saadi järgmised kaotatud kilogrammid:

Metoodika A	Metoodika B
10	18
22	12
15	16
9	8
18	20
11	17
6	15
12	14

Kas saab Wilcoxoni testi abil tõestada, et üks metoodika on tõhusam kui teine? Olgu olulisuse nivoo $\beta = 0.05$.

Ülesanne 3.20. Vantaa lennujaamas (Helsingi) registreeriti järgmised lennukite hilinemisajad:

Hilinemine minutites	Lendude arv
[0;20)	73
[20;40)	40
[40;60)	23
[60;80)	17
[80;100)	20
[100;120)	14
[120;140)	7
[140;160]	6

Kontrollige Kolmogorov-Smirnovi testiga, kas lennukite hilinemisaeg allub eksponentjaotusele. Olgu olulisuse nivoo $\beta = 0.05$.

Ülesanne 3.21. Genereerige 100 valimit $(X_1, X_2, ..., X_{20})$, mis esindavad normaaljaotust $\mathcal{N}(\theta, 1)$, kus $\theta = 1$.

1) Leidke igale valimile *bootstrap*- ja *jacknife*-hinnangud statistiku $\widehat{\theta} = \overline{X}$ dispersioonile. Leidke siis keskväärtuse ja standardhälbe hinnangud genereeritud 100 valimi põhjal.

2) Korrake punktis 1 tehtut statistikule $\widehat{\theta} = \overline{X}^2$. Võrrelge saadud tulemusi.

Ülesanne 3.22. Veenduge, et $\widehat{\theta} = \overline{x}$ korral on meetodil *jacknife*-meetodil leitud standardvea hinnang

$$s_{viga}^{jack} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n(n-1)}}$$

on nihketa.

Ülesanne 3.23. Meid huvitava parameetri hindamiseks koostatakse talle korrigeeritud nihkega 0.95-usaldusintervall. Saadakse järgmised asümmeetria- ja nihkeparandid: $\widehat{a} = 0.092$ ning $\widehat{b} = 0.185$. Leidke seoste (3.10)-(3.11) põhjal tõenäosused α_1 ning α_2.

Ülesanne 3.24. Olgu $m = C_n^k$. Tõestage, et siis

$$P\left(\text{ASL} = \frac{l}{m}\right) = \frac{1}{m}, \quad l = 1, 2, ..., m.$$

Ülesanne 3.25. Mõõdeti kahte erinevat tüüpi patareide eluigasid päevades. Saadi järgmised tulemused:

1) I tüüp: $\{342, 401, 315, 389, 398, 366\}$;

2) II tüüp: $\{402, 205, 311, 308, 395\}$.

Kas võib nende tulemuste põhjal kummutada hüpoteesi, et mõlemad patareid peavad vastu võrdse aja? Olulisuse nivooks võtta 0.1.

4. peatükk

Juhuslik vektor ja Bayesi statistika

Selles peatükis käsitleme Bayesi statistikat, mis uurib tegelikkust ühenduses teoreetilise eelteadmiste ja empiirikaga. Bayesi statistika sisaldab endas nii tõenäosusteooriat kui ka matemaatilist statistikat. Enne Bayesi statistika juurde asumist anname ülevaate mitmemõõtmeliste jaotuste teooriast. Bayesi meetodite rakendamine statistikas põhineb suuresti selle teooria tulemustel. Käsitleme selliseid mõisteid nagu juhusliku vektori ühisjaotus, tema tinglik jaotus, tinglik keskväärtus ja regressioonikordaja. Sissejuhatuseks defineerime mõiste juhuslik vektor. See on mitmemõõtmelise statistika alusmõiste.

Definitsioon 4.1. Vektorit, mille komponentideks on juhuslikud suurused, nimetatakse juhuslikuks vektoriks.

Juhuslik vektor on vektor algebralises mõttes. See tähendab, et kehtivad samad liitmise ning skalaariga korrutamise reeglid mis vektorruumi definitsioonis. Tähistagem juhuslikku vektorit kui

$$\mathbf{X} = (X_1, X_2, ..., X_k)^\intercal.$$

Seega on tegemist k-komponendilise veeruvektoriga.

4.1. Mitmemõõtmelise statistika alused

See osa keskendub mitmemõõtmelise statistikale, mille uurimisobjekt on juhuslik vektor, mida käsitleme eraldi diskreetsel ja pideval juhul.

4.1.1. Diskreetne juhuslik vektor

Uurime diskreetset juhuslikku vektorit sagedustabeli põhjal. Olgu meil juhuslik vektor $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$, kus $X = x_1, x_2, ..., x_k$ ja $Y = y_1, y_2, ..., y_l$. Juhusliku vektori $\mathbf{X}$ jaotusseadust kirjeldab kahemõõtmeline sagedustabel

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_l	X jaotus
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1l}	$p_{1.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{il}	$p_{i.}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$	$\vdots$
x_k	p_{k1}	$\cdots$	p_{kj}	$\cdots$	p_{kl}	$p_{k.}$
Y jaotus	$p_{.1}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.l}$	$\Sigma = 1$

Selles sagedustabelis ühistõenäosus

$$p_{ij} = P(X = x_i, Y = y_j)$$

ning komponentide üksiktõenäosused

$$p_{i.} = \sum_{j=1}^{l} p_{ij} = P(X = x_i)$$

ja

$$p_{.j} = \sum_{i=1}^{k} p_{ij} = P(Y = y_j).$$

Kehtib seaduspära

$$\Sigma = \sum_{j=1}^{l} p_{.j} = \sum_{i=1}^{k} p_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} = 1.$$

Ühistõenäosuse ja üksiktõenäosuste kaudu saab defineerida juhuslike suuruste sõltumatuse.

Definitsioon 4.2. Juhuslikud suurused X ja Y on sõltumatud parajasti siis, kui

$$p_{ij} = p_{i.}p_{.j}.$$

Seostega

$$E(X^m) = \sum_{i=1}^{k} x_i^m p_{i.} \text{ ning } E(Y^l) = \sum_{j=1}^{l} y_j^m p_{.j}$$

leitakse m-järku momendid komponentidele X ja Y. Nende komponentide m, n-segamoment

$$E(X^m Y^n) = \sum_{i=1}^{k} \sum_{j=1}^{l} x_i^m y_j^m p_{ij}.$$

Seega sõltub juhuslike suuruste X ja Y vaheline kovariatsioon nende komponentide 1,1-segamomendist.

Uurime järgnevalt juhusliku vektori $\mathbf{X}$ komponentide X ja Y vahelist regressioonikordajat. Tegemist on ühepoolse seosekordajaga, mille leidmine põhineb tinglikul keskväärtusel.

Definitsioon 4.3. Olgu juhuslikud suurused X ja Y juhusliku vektori $\mathbf{X}$ komponendid. Siis juhusliku suuruse X tinglik keskväärtus tingimusel Y

$$E(X|Y) = E(X|Y = y_j) = \sum_{i=1}^{k} x_i P(X = x_i|Y = y_j) =$$

$$= \sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{.j}}$$

ning juhusliku suuruse Y tinglik keskväärtus tingimusel X

$$E(Y|X) = E(Y|X = x_i) = \sum_{j=1}^{l} y_j P(Y = y_j|X = x_i) =$$

$$= \sum_{j=1}^{l} y_j \frac{p_{ij}}{p_{i.}}$$

Suhteid $\frac{p_{ij}}{p_{.j}}$ ning $\frac{p_{ij}}{p_{i.}}$ nimetatakse tinglikeks tõenäosusteks. Eeskiri, kuidas leida tinglikke keskväärtusi sagedustabeli abil, pannakse kirja järgmiselt:

$$E(X \mid Y = y_j) = \sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{.j}} \tag{4.1}$$

ning

$$E(Y \mid X = x_i) = \sum_{j=1}^{l} y_j \frac{p_{ij}}{p_{i.}}. \tag{4.2}$$

Tinglik keskväärtus iseloomustab komponentide X ja Y vahelist keskmist sõltuvust regressiooni mõttes.

Tõestame järgnevalt tingliku keskväärtuse mõningad omadused.

Lause 4.1. Tinglikul keskväärtusel on järgmised omadused.

1° Kui juhuslikud suurused X ja Y on sõltumatud,

$$E(X|Y) = E(X) \text{ ja } E(Y|X) = E(Y).$$

2° Keskväärtus tinglikust keskväärtusest on keskväärtus ehk

$$E(E(X|Y)) = E(X) \text{ ja } E(E(Y|X)) = E(Y).$$

3° Kehtivad seosed

$$D(E(X|Y)) \leq D(X) \text{ ja } D(E(Y|X)) \leq D(Y).$$

Tõestus. Esmalt tõestame esimese omaduse. Kuna juhuslikud suurused X ja Y on sõltumatud, siis definitsiooni 4.2 kohaselt

$$p_{ij} = p_{i.}p_{.j}.$$

Seega

$$E(X|Y) = \sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{.j}} = \sum_{i=1}^{k} x_i \frac{p_{i.}p_{.j}}{p_{.j}} = \sum_{i=1}^{k} x_i p_{i.} = E(X)$$

ning

$$E(Y|X) = \sum_{j=1}^{l} y_j \frac{p_{ij}}{p_{i.}} = \sum_{j=1}^{l} y_j \frac{p_{i.}p_{.j}}{p_{i.}} = \sum_{j=1}^{l} y_j p_{.j} = E(Y).$$

Teise omaduse puhul saame, et

$$E(E(X|Y)) = \sum_{j=1}^{l}\sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{.j}} p_{.j} =$$

$$= \sum_{i=1}^{k} x_i \sum_{j=1}^{l} p_{ij} = \sum_{i=1}^{k} x_i p_{i.} = E(X)$$

ning

$$E(E(Y|X)) = \sum_{i=1}^{k}\sum_{j=1}^{l} y_j \frac{p_{ij}}{p_{i.}} p_{i.} =$$

$$= \sum_{j=1}^{l} y_j \sum_{i=1}^{k} p_{ij} = \sum_{j=1}^{l} y_j p_{.j} = E(Y).$$

Kolmanda omaduse tõestamisel lähtume seostest

$$D(E(X|Y)) = E(E^2(X|Y)) - E^2(X)$$

ja

$$D(X) = E(X^2) - E^2(X).$$

Meil tuleb näidata, et $E(E^2(X|Y)) \leq E(X^2)$. Saame, et

$$E(E^2(X|Y)) = \sum_{j=1}^{l} \left(\sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{.j}} \right)^2 p_{.j} \leq \sum_{j=1}^{l}\sum_{i=1}^{k} x_i^2 \frac{p_{ij}}{p_{.j}} p_{.j} =$$

$$\sum_{i=1}^{k} x_i^2 \sum_{j=1}^{l} p_{ij} = \sum_{i=1}^{k} x_i^2 p_{i.} = E(X^2).$$

Analoogiliselt saame tõestada, et $D(E(Y|X)) \leq D(Y)$.

□

Defineerime ühepoolsed regressioonikordajad kui

$$\gamma(X,Y) = \sqrt{\frac{D(E(X|Y))}{D(X)}}$$

ja

$$\gamma(Y,X) = \sqrt{\frac{D(E(Y|X))}{D(Y)}}.$$

Tingliku keskväärtuse omadusest 3° järeldub, et $0 \leq \gamma(X,Y) \leq 1$ ja $0 \leq \gamma(Y,X) \leq 1$. Saadud seosekordajad on tõesti ühepoolsed, sest üldjuhul

$$\gamma(X,Y) \neq \gamma(Y,X).$$

Kui juhuslikud suurused X ja Y on sõltumatud, siis

$$\gamma(X,Y) = \gamma(Y,X) = 0.$$

Kui aga juhuslik suurus Y on esitatav juhusliku suuruse X ühese funktsioonina, siis

$$\gamma(X,Y) = \gamma(Y,X) = 1.$$

Regressioonikordaja on invariantne lineaarteisenduse suhtes, see tähendab

$$\gamma(aY+b,X) = \gamma(Y,X), \quad a \neq 0.$$

Regressioonikordajate avaldises olevad tingliku keskväärtuse dispersioonid leitakse järgmiselt:

$$D(E(X|Y)) = E(E^2(X \mid Y)) - E^2(E(X \mid Y)) =$$

$$= \sum_{j=1}^{l} \left(\sum_{i=1}^{k} x_i \frac{p_{ij}}{p_{\cdot j}} \right)^2 p_{\cdot j} - E^2(X)$$

ning

$$D(E(Y|X)) = E(E^2(Y \mid X)) - E^2(E(Y \mid X)) =$$

$$= \sum_{i=1}^{k} \left(\sum_{j=1}^{l} y_i \frac{p_{ij}}{p_{i\cdot}} \right)^2 p_{i\cdot} - E^2(Y).$$

Toome näite juhuslikust vektorist $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$, kus $\gamma(X,Y) = 0$, aga $\gamma(Y,X) = 1$.

Näide 4.1. Olgu meil juhuslik vektor $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$, mille jaotust kirjeldab järgmine sagedustabel:

x_i \ y_j	0	1	X jaotus
-1	0	$\frac{1}{4}$	$\frac{1}{4}$
0	$\frac{1}{2}$	0	$\frac{1}{2}$
1	0	$\frac{1}{4}$	$\frac{1}{4}$
Y jaotus	$\frac{1}{2}$	$\frac{1}{2}$	$\Sigma = 1$

Selle tabeli põhjal saame, et keskväärtused

$$E(X) = 0 \text{ ja } E(Y) = \frac{1}{2}$$

ning dispersioonid

$$D(X) = \frac{1}{2} \text{ ja } D(Y) = \frac{1}{4}.$$

Seoste (4.1)-(4.2) põhjal saame tinglikeks keskväärtusteks

$$E(X|Y) = \begin{cases} 0, \text{ kui } Y = 0, \\ 0, \text{ kui } Y = 1, \end{cases}$$

ja

$$E(Y|X) = \begin{cases} 1, \text{ kui } X = -1, \\ 0, \text{ kui } X = 0, \\ 1, \text{ kui } X = 1. \end{cases}$$

Leiame saadud tinglike keskväärtuste dispersioonid. Saame, et

$$D(E(X \mid Y) = D(E(X)) = 0$$

ja

$$D(E(Y|X)) = E(E^2(Y|X)) - E^2(Y) =$$

$$= 1 \cdot P(X = -1) + 0 \cdot P(X = 0) + 1 \cdot P(X = 1) - \frac{1}{4} = \frac{1}{4} + \frac{1}{4} - \frac{1}{4} = \frac{1}{4} = D(Y).$$

Seega oleme saanud järgmised regressioonikordajad:

$$\gamma(X,Y) = 0 \ \text{ ja } \ \gamma(Y,X) = 1.$$

Antud juhul võib öelda, et juhuslik suurus Y sõltub juhuslikust suurusest X, kuid juhuslik suurus X ei sõltu juhuslikust suurusest Y.

Kõige üldisema seose komponentide X ja Y vahel annab Tšuprovi seosekordaja ehk χ^2-kordaja. Selle kordaja tarbeks tuuakse sisse suurus

$$H(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}.$$

On ilmne, et $H(X,Y) \geq 0$. Mis aga on $H(X,Y)$ ülemine piir? Selle leidmiseks defineerime suuruse

$$H_0(k,l) = \min\{k-1; l-1\}.$$

Lause 4.2. Kehtib seos

$$H(X,Y) \leq H_0(k,l).$$

Tõestus. Saame, et

$$H(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{p_{ij}^2 - 2p_{ij}p_{i.}p_{.j} + p_{i.}^2p_{.j}^2}{p_{i.}p_{.j}}\right) = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{p_{ij}^2}{p_{i.}p_{.j}} - 2p_{ij} + p_{i.}p_{.j}\right).$$

On ilmne, et

$$p_{ij} \leq p_{i.}.$$

Seega

$$\sum_{i=1}^{k}\frac{p_{ij}^2}{p_{i.}p_{.j}} \leq \frac{1}{p_{.j}}\sum_{i=1}^{k}p_{ij} = \frac{p_{.j}}{p_{.j}} = 1.$$

Järelikult

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{p_{ij}^2}{p_{i.}p_{.j}} - 2p_{ij} + p_{i.}p_{.j}\right) \le \sum_{j=1}^{l}\left((1 - 2\sum_{i=1}^{k} p_{ij} + p_{.j}\sum_{i=1}^{k} p_{i.}\right) =$$

$$= \sum_{j=1}^{l}(1 - p_{.j}) = l - 1.$$

Teises järjekorras summeerides saame, et $H(X,Y) \le k - 1$. Seega

$$H(X,Y) \le H_0(k,l).$$

□

Tõestatud lausest järeldub, et

$$\frac{H^2(X,Y)}{(k-1)(l-1)} \le 1. \tag{4.3}$$

Tuginedes sellele järeldusele ja lausele 4.2, defineeritake Tšuprovi seosekordaja (või χ^2-kordaja) kui

$$\chi(X,Y) = \sqrt{\frac{H(X,Y)}{\sqrt{(k-1)(l-1)}}}.$$

Tšuprovi seosekordaja iseloomustab komponentide X ja Y vahelist statistilist sõltuvust. Seda kordajat saab interpreteerida järgmiselt.

1) Kui juhuslikud suurused X ja Y on sõltumatud, siis $\sum_{i=1}^{k}\sum_{j=1}^{l}(p_{ij} - p_{i.}p_{.j})^2 = 0$. Seega $\chi(X,Y) = 0$.

2) Kui juhuslikud suurused X ja Y on üksüheses sõltuvuses, siis $p_{i.} = p_{ij} = p_{.j}$ (miks?) ning seoses (4.3) on võrratuse asemel võrdus. Seega $\chi(X,Y) = 1$.

3) Kui juhuslike suuruste X ja Y vahel on funktsionaalne seos, mis ei pruugi olla üksühene, siis

$$\chi(X,Y) = \sqrt[4]{\frac{l'-1}{k'-1}},$$

kus $l' = \min\{k; l\}$ ja $k' = \max\{k; l\}$.

Vaatame näidet χ^2-seosekordaja leidmisest juhul, kui juhuslike suuruste X ja Y vahel esineb funktsionaalne seos.

Näide 4.2. Olgu juhuslik suurus $X = \{-1, 0, 1\}$ selline, et

$$P(X = -1) = 0.2,$$

$$P(X = 0) = 0.4$$

ja

$$P(X = 1) = 0.4.$$

Olgu $Y = X^2$

Siis juhusliku vektori $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ jaotusseadus on järgmine:

x_i \ y_j	0	1	X jaotus
-1	0	0.2	0.2
0	0.4	0	0.4
1	0	0.4	0.4
Y jaotus	0.4	0.6	$\Sigma = 1$

Saame, et $H(X, Y) = 1$ ja $\chi(X, Y) \approx 0.841$. Antud juhul $\min\{k; l\} = 2$ ja $\max\{k; l\} = 3$ ning $\sqrt[4]{\frac{1}{2}} \approx 0.841$.

Alljärgnevas näites uurime komponentide X ja Y vahelist lineaarset korrelatsiooni, regressioonisõltuvust ning leiame χ^2-seosekordaja.

Näide 4.3. Tuleme tagasi näite 1.21 juurde. Selles näites testiti filmi meeldimise sõltuvust vanusest. Vanust iseloomustas seal tunnus $X = 0, 1, 2$ ning filmi meeldimist tunnus $Y = -1, 0, 1$. Mõlemat tunnust võib vaadelda kui järjestatavat. Näites 1.21 toodud küsitlustulemuste põhjal saame järgmise sagedustabeli:

x_i \ y_j	-1	0	1	$p_{i.}$
0	0.045	0.082	0.2	0.327
1	0.1	0.045	0.155	0.3
2	0.155	0.136	0.082	0.373
$p_{.j}$	0.3	0.264	0.436	$\sum = 1$

1) Leiame komponentide X ja Y vahelise lineaarse korrelatsioonikordaja. Momentide leidmise eeskirja põhjal saame, et

$$E(X) = \sum_{i=1}^{3} x_i p_{i.} = 1.045 \text{ ja } E(Y) = \sum_{j=1}^{3} y_j p_{.j} \approx 0.136$$

ning

$$E(X^2) = \sum_{i=1}^{3} x_i^2 p_{i.} \approx 1.79 \text{ ja } E(Y^2) = \sum_{j=1}^{3} y_j^2 p_{.j} = 0.736.$$

Komponentide X ja Y dispersioonideks saame, et

$$D(X) = E(X^2) - E^(X) = 0.698 \text{ ja } D(Y) = E(Y^2) - E^(Y) = 0.718.$$

Nende komponentide 1,1-segamoment

$$E(XY) = \sum_{i=1}^{3} \sum_{j=1}^{3} x_i y_j p_{ij} \approx -0.091.$$

Tunnuste X ja Y vaheline kovariatsioon

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y) \approx -0.233.$$

Seega on vanus ja filmi meeldimine negatiivselt korreleeritud. Seega, mida noorem on vaataja, seda rohkem film meeldib. Kui tugev aga on see korrelatsioon? Sellele annab vastuse lineaarne korrelatsioonikordaja

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \approx -0.33.$$

Järelikult on tegemist keskmise negatiivse korrelatsiooniga. Kumb sõltuvus aga on tugevam, kas vanuse sõltuvus filmi meeldivusest või filmi meeldivuse sõltuvus vanusest? Seda uurime järgmises punktis.

2) Leiame ühepoolsed regressioonikordajad. Esmalt leiame tinglikud keskväärtused $E(X \mid Y)$ ning $E(Y \mid X)$. Saame, et

$$E(X|Y) = \sum_{i=1}^{3} x_i \frac{p_{ij}}{p_{.j}} = \begin{cases} 1.364, \text{ kui } Y = -1, \\ 1.207, \text{ kui } Y = 0, \\ 0.729, \text{ kui } Y = 1 \end{cases}$$

ja

$$E(Y|X) = \sum_{j=1}^{3} y_j \frac{p_{ij}}{p_{i.}} = \begin{cases} 0.472, \text{ kui } X = 0, \\ 0.182, \text{ kui } X = 1, \\ -0.195, \text{ kui } X = 2. \end{cases}$$

Lugeja saab ise kontrollida tingliku keskväärtuse omaduse 2° kehtivust. Edasi tuleb leida dispersioonid

$$D(E(X \mid Y)) \text{ ning } D(E(Y \mid X)).$$

Saame järgmised tulemused:

$$D(E(X \mid Y)) = E(E^2(X \mid Y)) - E^2(X) \approx 1.151$$

ning

$$D(E(Y \mid X)) = E(E^2(Y \mid X)) - E^2(X) \approx 0.057.$$

Seega saame regressioonikordajateks

$$\gamma(X, Y) = \sqrt{\frac{D(E(X \mid Y))}{D(X)}} \approx 0.465$$

ning

$$\gamma(Y, X) = \sqrt{\frac{D(E(Y \mid X))}{D(Y)}} \approx 0.282.$$

Seega on vanuse sõltuvus filmi meeldimisest tugevam kui filmi meeldimise sõltuvus vanusest.

3) Leiame χ^2-seosekordaja. Lõpuks uurime kõige üldisemat seost tunnuste X ja Y vahel. Koostame sagedustabeli eeldusel, et X ja Y on sõltumatud. Definitsiooni 4.2 põhjal saame siis tabelile järgmise kuju:

x_i \ y_j	-1	0	1
0	0.098	0.086	0.143
1	0.09	0.079	0.131
2	0.112	0.098	0.163

Suurus

$$H(X,Y)=\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}\approx 0.142$$

Seega saame antud juhul χ^2-seosekordajaks

$$\chi(X,Y)\sqrt{\frac{H(X,Y)}{\sqrt{(3-1)(3-1)}}}\approx 0.267.$$

4.1.2. Pidev juhuslik vektor

Olgu meil juhuslik vektor $\mathbf{X}=(X_1,X_2,...,X_k)^\intercal$ pidevate komponentidega, see tähendab juhuslikud suurused X_i, $i=1,2,...,n$ on pidevad.

Definitsioon 4.4. Funktsiooni

$$F(x_1,x_2,...,x_k)=P(X_1\le x_1,X_2\le x_2,...,X_k\le x_k)$$

nimetatakse juhusliku vektori $\mathbf{X}=(X_1,X_2,...,X_k)^\intercal$ jaotusfunktsiooniks.

Pideva juhusliku vektori $\mathbf{X}$ tihedusfunktsioon

$$f(x_1,x_2,...,x_k)=\frac{\partial^k F(x_1,x_2,...,x_k)}{\partial x_1\partial x_2...\partial x_k}$$

ehk

$$F(x_1,x_2,...,x_k)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_k}f(u_1,u_2,...,u_k)du_1du_2...du_k.$$

Uurime lähemalt juhtu, kui $k=2$. Olgu meil juhuslik vektor $\mathbf{X}=(X,Y)^\intercal$. Selle juhusliku vektori tihedusfunktsioon

$$f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y}.$$

Komponendi X tihedusfunktsioon

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

ning komponendi Y tihedusfunktsioon

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Tihedusfunktsiooni $f(x,y)$ nimetatakse juhusliku vektori $\mathbf{X}$ komponetide ühistiheduseks, tihedusfunktsioone $f_1(x)$ ja $f_2(y)$ aga tema komponentide üksiktihedusteks ehk marginaalseteks tihedusteks.

Lause 4.3. Juhusliku vektori $\mathbf{X}$ komponendid X ja Y on sõltumatud parajasti siis, kui

$$f(x,y) = f_1(x)f_2(y).$$

Uurime tinglikke keskväärtusi $E(X \mid Y)$ ja $E(Y \mid X)$ pideval juhul. Selleks toome sisse mõiste tinglik jaotustihedus. Lähtume matemaatilisest analüüsist tuntud kahe muutuja funktsiooni tuletise definitsioonist:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \lim_{(\Delta x, \Delta y)\to(0,0)} \frac{F(x+\Delta x, y+\Delta y) - F(x,y)}{\Delta x \Delta y}.$$

Juhusliku vektori tihedusfunktsiooni puhul avaldub see definitsioon kui

$$f(x,y) = \lim_{\Delta y \to 0} \lim_{\Delta x \to 0} \frac{P((x \le X \le x+\Delta x)(y \le Y \le y+\Delta y))}{\Delta x \Delta y} =$$

$$= \lim_{y\to 0} \lim_{x \to 0} \frac{P(x \le X \le x+\Delta x)}{\Delta x} \frac{P(y \le Y \le y+\Delta y \mid x \le X \le x+\Delta x)}{\Delta y} =$$

$$= \lim_{y\to 0} f_1(x) \frac{P(y \le Y \le y+\Delta y \mid X = x)}{\Delta y} = f_1(x) f_2(y \mid x).$$

Saadu põhjal võime kirja panna järgmise definitsiooni.

Definitsioon 4.5. Funktsiooni

$$f_2(y \mid x) = \lim_{\Delta y \to 0} = \frac{P(y \le Y \le y+\Delta y \mid X = x)}{\Delta y}$$

nimetatakse juhusliku vektori $(X,Y)^\intercal$ komponendi Y tinglikuks jaotustiheduseks komponendi X suhtes.

Analoogiliselt saame defineerida komponendi X tingliku jaotustiheduse komponendi Y suhtes

$$f_1(x \mid y) = \lim_{\Delta x \to 0} = \frac{P(x \leq X \leq x + \Delta x \mid Y = y)}{\Delta x}.$$

Lihtne on veenduda, et

$$f_2(y \mid x) = \frac{f(x,y)}{f_1(x)} \text{ ja } f_1(x \mid y) = \frac{f(x,y)}{f_2(y)}.$$

Üldistame tingliku tiheduse k-mõõtmelisele juhule, defineerides täistingliku jaotuse. Selleks toome sisse vektori $\mathbf{x}_{-i}$, mis on saadud pärast i-nda komponendi elimineerimist k-mõõtmelisest vektorist $\mathbf{x}$.

Definitsioon 4.6. Jaotust nimetatakse täistinglikuks jaotuseks, kui seda kirjeldab tinglik tihedusfunktsioon $f_i(x_i \mid \mathbf{x}_{-i})$, $i = 1, 2, ..., k$.

Uurime järgnevalt tinglikke keskväärtusi pideval juhul. Tinglike jaotustiheduste kaudu saame, et

$$E(Y \mid X) = \int_{-\infty}^{\infty} y f_2(y \mid x) dy$$

ning

$$E(X \mid Y) = \int_{-\infty}^{\infty} x f_1(x \mid y) dx.$$

Tõestame nüüd tingliku keskväärtuse omaduse 2° pideval juhul. Saame, et

$$E(E(Y \mid X)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_2(y \mid x) dy f_1(x) dx =$$

$$= \int_{-\infty}^{\infty} y dy \int_{-\infty}^{\infty} f(x,y) dx = \int_{-\infty}^{\infty} y f_2(y) dy = E(Y)$$

ning

$$E(E(X \mid Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_2(x \mid y) dx f_2(y) dy =$$

$$= \int_{-\infty}^{\infty} x dx \int_{-\infty}^{\infty} f(x,y) dy = \int_{-\infty}^{\infty} x f_1(x) dx = E(X).$$

Tinglikud keskväärtused kirjeldavad komponentide X ja Y vahelist keskmist regressioonisõltuvust. Olgu komponendi Y keskmine sõltuvus komponendist X määratud funktsiooniga g_1 ja komponendi X keskmine sõltuvus komponendist Y määratud funktsiooniga g_2. Siis

$$y = g_1(x) = E(Y \mid X = x) \text{ ning } x = g_2(y) = E(X \mid Y = y).$$

Lihtne on veenduda, et

$$E(Y) = \int_{-\infty}^{\infty} g_1(x) f_1(x) dx \text{ ning } E(X) = \int_{-\infty}^{\infty} g_2(y) f_2(y) dy.$$

Regressioonikordajate $\gamma(X, Y)$ ning $\gamma(Y, X)$ arvutamiseks tuleb meil leida integraalid

$$E(E^2(X \mid Y)) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g_1(x) f_1(x) dx \right)^2 f_2(y) dy$$

ning

$$E(E^2(Y \mid X) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g_2(y) f_2(y) dy \right)^2 f_1(x) dx.$$

4.1.3. Tinglik keskväärtus ja regressioonanalüüs

Uurime regressioonanalüüsi tingliku keskväärtuse kontekstis. Olgu meil uuritav tunnus Y ja faktortunnused $X_1, X_2, ..., X_k$. Olgu $x_1, x_2, ..., x_k$ nende faktortunnuste mõõdetud väärtused. Toome sisse funktsiooni

$$h(\mathbf{x}) = h(x_1, x_2, ..., x_k) = E(Y | X_1 = x_1, X_2 = x_2, ..., X_k = x_k) =$$

$$= E(Y | \mathbf{X} = \mathbf{x}).$$

Definitsioon 4.7. Funktsiooni h nimetatakse tunnuse Y regressioonifunktsiooniks tingimusel $\mathbf{X} = \mathbf{x}$.

Erijuhul, kui $k = 1$, saame, et $h(x) = E(Y \mid X = x)$ ehk sel juhul võrdub regressioonifunktsioon juhusliku suuruse Y tingliku keskväärtusega tingimusel X.

Defineerime funktsiooni, mida tähistame kui $d(\mathbf{X})$. Nimetagem seda juhusliku suuruse Y ennustajaks (ingl *predictor*) juhusliku vektori $\mathbf{X}$ kaudu. Üldise lineaarse mudeli korral

$$d(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k.$$

Olgu ennustaja $d(\mathbf{X})$ ruutviga $E(Y - d(\mathbf{X}))^2$.

Definitsioon 4.8. Öeldakse, et ennustaja d_2 on parem ennustajast d_1, kui

$$E(Y - d_2(\mathbf{X}))^2 \leq E(Y - d_1(\mathbf{X}))^2.$$

Vaatame nüüd lähemalt juhtu, kus $k = 1$. Sel juhul on ennustaja funktsioon tunnusest X ning selle ruutveaks on keskväärtus $E(Y - d(Y))^2$. Kui ennustaja on lineaarne, siis

$$d(X) = \beta_0 + \beta_1 X$$

ning ruutveaks on $E(Y - \beta_0 - \beta_1 X)^2$. Millised aga oleksid kordajad β_0 ja β_1, mis minimeeriksid antud ruutvea? Sellele küsimusele annab vastuse alljärgnev teoreem.

Teoreem 4.1. Eeldame, et $E(X^2) < \infty$ ning $E(Y^2) < \infty$. Olgu $\mu_x = E(X)$, $\mu_y = E(Y)$, $\sigma_x^2 = D(X)$, $\sigma_y^2 = D(Y)$, $\sigma_{xy} = \text{cov}(X, Y)$ ning $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$. Siis parim lineaarne ennustaja

$$Y = L(X) = \beta_0 + \beta_1 X,$$

kus

$$\beta_0 = \mu_y - \frac{\sigma_{xy}}{\sigma_x^2}\mu_x \text{ ning } \beta_1 = \rho\frac{\sigma_y}{\sigma_x}.$$

Pärast mõningaid teisendusi saame teoreemi 4.1 põhjal, et parim lineaarne ennustaja

$$L(X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x).$$

Milline on aga saadud parima lineaarse ennustaja ruutviga? Vastuse sellele küsimusele saab alljärgneva teoreemiga.

Teoreem 4.2. Parimale lineaarsele ennustajale vastav keskmine ruutviga

$$E(Y - L(X))^2 = \sigma_y^2(1 - \rho^2).$$

Tõestus. Saame, et

$$E(Y - L(X))^2 = E(Y - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x))^2 =$$

$$= E(Y - \mu_y)^2 + \rho^2\frac{\sigma_y^2}{\sigma_x^2}E(X - \mu_x)^2 - 2\rho\frac{\sigma_y}{\sigma_x}E(X - \mu_x)(Y - \mu_y) =$$

$$= \sigma_y^2 + \rho^2\sigma_y^2 - 2\rho\frac{\sigma_y}{\sigma_x}\sigma_{xy} = \sigma_y^2 - \rho^2\sigma_y^2 = \sigma_y^2(1 - \rho^2).$$

□

Suurust $\sigma_y^2(1 - \rho^2)$ nimetatakse lineaarse mudeli (ennustaja) jäägist tingitud hajuvuseks, mida tähistasime andmeanalüüsi peatüki üldiste lineaarsete mudelite osas kui S_{res}. Kui $\mid \rho \mid = 1$, siis

$$S_{res} = E(Y - L(X))^2 = 0.$$

Kui aga $\mid \rho \mid = 0$, siis

$$S_{res} = E(Y - L(X))^2 = \sigma_y^2.$$

Lihtne on veenduda, et uuritava tunnuse Y hajuvus

$$\sigma_y^2 = \sigma_y^2\rho^2 + \sigma_y^2(1 - \rho^2).$$

Soovitatav on lugejal võrrelda seda seost lausega 2.2 osas 2.2.2.

Demonstreerime lineaarse ennustaja leidmist järgmise näitega.

Näide 4.4. Olgu meil juhuslik vektor $\mathbf{X} = (X, Y)^\intercal$ ühistihedusega

$$f(x, y) = \begin{cases} 8xy, \text{ kui } 0 < y < x < 1, \\ 0, \text{ mujal .} \end{cases}$$

Leiame parima lineaarse ennustaja $Y = L(X)$. Selleks leiame esmalt komponentide X ja Y üksiktihedused. Saame, et

$$f_1(x) = 8\int_0^x xydy = 4x^3 \text{ ja } f_2(y) = 8\int_y^1 xydx = 4y(1-y^2).$$

Ühistihedusest järeldub, et $f_1(x) = 0$, kui $x \notin [0;1]$, ning $f_2(y) = 0$, kui $y \notin [0;1]$. Komponentide X ja Y keskväärtused

$$E(X) = \int_0^1 4x^4dx = \frac{4}{5} \text{ ja } E(Y) = \int_0^1 4y^2(1-y^2)dy = \frac{8}{15}$$

ning dispersioonid

$$D(X) = \int_0^1 4x^5dx - \frac{16}{25} = \frac{2}{75} \text{ ja } D(Y) = \int_0^1 4y^3(1-y^2)dy - \frac{64}{225} = \frac{11}{225}.$$

Leidmaks komponentide X ja Y vahelist kovariatsiooni tuleb meil leida keskväärtus

$$E(XY) = 8\int_0^1\int_0^x x^2y^2dxdy = \frac{8}{3}\int_0^1 x^5dx = \frac{4}{9}.$$

Seega kovariatsioon

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{4}{9} - \frac{8}{15}\frac{4}{5} = \frac{4}{225}$$

ning korrelatsioon

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{2\sqrt{66}}{33}.$$

Pärast mõningaid teisendusi saame parimaks lineaarseks ennustajaks

$$Y = L(X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X-\mu_x) = \frac{4}{5} + \frac{4}{11}\left(X - \frac{8}{15}\right) = \frac{20}{33} + \frac{4}{11}X.$$

Seega on antud ühisjaotusele vastava parima regressioonisirge võrrandiks

$$y = \frac{20}{33} + \frac{4}{11}x.$$

4.2. Bayesi meetodid

Käsitleme ühte nüüdisaegse matemaatilise statistika haru, mida nimetatakse Bayesi statistikaks. Klassikaline statistika käsitleb jaotuse parameetrit kui konstantset suurust, mida hinnatakse valimi põhjal. Bayesi statistikas on aga jaotuse parameeter juhuslik suurus, mille jaotust korrigeeritakse empiirika abil. Tänapäeval leiab Bayesi statistika üha enam rakendust loodus- ja inseneriteaduste valdkonna probleemide lahendamisel.

4.2.1. Täistõenäosus ja Bayesi valem klassikalisel kujul

Tuletame meelde täistõenäosuse ja Bayesi valemid n-ö klassikalisel kujul ehk tõenäosusteooriast tuntud kujul. Need põhinevad tinglikul tõenäosusel. Defineerime esmalt ühe sündmuste süsteemi.

Definitsioon 4.9. Sündmuste süsteemi $\mathcal{A} = \{H_1, H_2, ..., H_m\}$ nimetatakse hüpoteesideks, kui

1° $\forall i = 1, 2, ..., m: \quad H_i \neq \emptyset$;

2° $\forall i, j = 1, 2, ..., m, \quad i \neq j: \quad H_i \cap H_j = \emptyset$;

3° $H_1 \cup H_2 \cup ... \cup H_m = \Omega$.

Sõltugu sündmuse A tõenäosus hüpoteesidest $H_1, H_2, ..., H_m$. Nendes tähistustes saame täistõenäosuse valemi

$$P(A) = \sum_{i=1}^{m} P(A|H_i)P(H_i)$$

ja Bayesi valemi

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{P(A)} \quad j = 1, 2, ..., m.$$

Näide 4.5. Allugu lambipirni eluiga T kuudes eksponentjaotusele, mille parameetriks olgu m_i, $i = 1, 2, 3, 4$. Olgu 40% lambipirnide keskmine

eluiga 8 kuud, 30% pirnidest pidagu keskmiselt vastu 10 kuud, 20% pirnidest 12 kuud ning 10% lambipirnide puhul olgu keskmine eluiga 18 kuud. Eesmärk on leida üle aasta vastu pidanud lambipirnide keskmiste eluigade protsentuaalne jaotus.

Kuna tegemist on eksponentjaotusega, siis keskväärtus $E(T) = \frac{1}{m_i}$. Antud juhul on parameetrite väärtused järgmised:

$$m_1 = \frac{1}{8}; \quad m_2 = \frac{1}{10}; \quad m_3 = \frac{1}{12}; \quad m_4 = \frac{1}{18}.$$

Juhusliku suuruse M jaotuseks saame, et

$$P(M = m_1) = 0.4; \quad P(M = m_2) = 0.3;$$
$$P(M = m_3) = 0.2; \quad P(M = m_4) = 0.1.$$

Seega täistõenäosus

$$P(T > t) = \sum_{i=1}^{4} P(T > t | M = m_i) P(M = m_i) =$$

$$= \exp\left(-\frac{t}{8}\right)0.4 + \exp\left(-\frac{t}{10}\right)0.3 + \exp\left(-\frac{t}{12}\right)0.2 + \exp\left(-\frac{t}{18}\right)0.1.$$

Kui $t = 12$, siis $P(T > t) \approx 0.305$. Bayesi valemit rakendades saame, et

$$P(M = m_i \mid T > 12) = \frac{P(T > 12 | M = m_i) P(M = m_i)}{P(T > 12)}.$$

Asendades m_i tema konkreetsete väärtustega, saame, et

$$P(M = \frac{1}{8} \mid T > 12) = 0.293, \quad P(M = \frac{1}{10} \mid T > 12) = 0.297,$$
$$P(M = \frac{1}{12} \mid T > 12) = 0.242 \text{ ning } P(M = \frac{1}{18} \mid T > 12) = 0.169.$$

Seega saime üle aasta vastu pidanud lambipirnidele järgmise protsentuaalse jaotuse: 29.3% lambipirnide keskmine eluiga on 8 kuud, 29.7% pirnidest on keskmise elueaga 10 kuud, 24.4% pirnidest on keskmise elueaga 12 kuud ning 16.9% lambipirnidest peavad keskmiselt vastu 18 kuud.

4.2.2. Tinglikustamise võte

Järgnevalt üldistame täistõenäosuse ja Bayesi valemid pidevale juhule. Olgu meil juhuslik suurus X tihedusfunktsiooniga f. Olgu θ selle funktsiooni parameeter, mis aga on omakorda juhuslik suurus tihedusfunktsiooniga p. Kui parameetri θ jaotus on pidev, siis avaldub täistõenäosuse valem järgmiselt:

$$f(x) = \int_{-\infty}^{\infty} f(x \mid \theta)p(\theta)d\theta. \tag{4.4}$$

Valemi (4.4) näol on meil tegemist tinglikustamise võtte eeskirjaga ehk tinglikustamisega (ingl *conditioning*). Demonstreerime seda valemit kahe näitega.

Näide 4.6. Olgu meil juhuslik suurus X Poissoni jaotusega parameetriga m. See parameeter olgu eksponentjaotusega juhusliku suuruse M mingi realisatsioon. Olgu keskväärtus $E(M) = 1$. Siis parameetri m tihedusfunktsioon

$$f_M(x) = \begin{cases} 0, \text{ kui } x < 0, \\ e^{-x}, \text{ kui } x \geq 0. \end{cases}$$

Seose (4.4) põhjal saame, et

$$P(X = k) = \int_0^{\infty} P(X = k|M = x)f_M(x)dx =$$

$$= \int_0^{\infty} e^{-x}\frac{x^k}{k!}e^{-x}dx = \int_0^{\infty} \frac{x^k}{k!}e^{-2x}dx.$$

Toome täistõenäosuse avaldisse Euleri gammafunktsiooni

$$\Gamma(k+1) = \int_0^{\infty} (2x)^{k+1-1}e^{-2x}d2x.$$

Gammafunktsiooni kohta on teada, et

$$\Gamma(k+1) = k!.$$

Seega võime teisendada

$$P(X=k)=\frac{1}{2^{k+1}}\int_0^\infty \frac{1}{\Gamma(k+1)}2^{k+1}x^{k+1-1}e^{-2x}dx=\frac{1}{2^{k+1}}=\frac{1}{2}\frac{1}{2^k}.$$

Saime, et juhuslik suurus X allub geomeetrilisele jaotusele, mille parameeter on $\frac{1}{2}$.

Näide 4.7. Vaatleme nüüd juhtu, kus juhuslik suurus X allub binoomjaotusele $B(n,p)$. Parameeter n aga olgu Poissoni jaotusega juhusliku suuruse N mingi realisatsioon. Seega vaatame olukorda, kus sõltumatute katsete hulk $N\sim Po(\lambda)$ ning õnnestumise tõenäosus on p. Järelikult keskväärtus $E(N)=\lambda$. Saame, et täistõenäosus

$$P(X=k)=\sum_{n=0}^{\infty}P(X=k|N=n)P(N=n)=\sum_{n=k}^{\infty}C_n^k p^k q^{n-k}e^{-\lambda}\frac{\lambda^n}{n!}=$$

$$=\frac{p^k}{k!}e^{-\lambda}\sum_{n=k}^{\infty}\frac{\lambda^n}{(n-k)!}q^{n-k}=\frac{(\lambda p)^k}{k!}e^{-\lambda}\sum_{n=k}^{\infty}\frac{(\lambda q)^{n-k}}{(n-k)!}=$$

$$=\frac{(\lambda p)^k}{k!}e^{-\lambda}\sum_{j=0}^{\infty}\frac{(\lambda q)^j}{j!}=\frac{(\lambda p)^k}{k!}e^{-\lambda}e^{\lambda q}=\frac{(\lambda p)^k}{k!}e^{\lambda(q-1)}=\frac{(\lambda p)^k}{k!}e^{-\lambda p}.$$

Seega allub juhuslik suurus X Poissoni jaotusele parameetriga λp.

4.2.3. Bayesi valem pideval juhul

Tõenäosusteooria ülesannetes eeldatakse, et meid huvitav juhuslik suurus X on jaotusega, mille parameetrid on kindlalt teada. Näiteks $X\sim Po(\lambda)$, $X\sim\mathcal{N}(\mu,\sigma)$ või siis, et münt ja täring on sümmeetriline.

Matemaatilise statistika mõttes aga on jaotuse parameetrid tundmatud ning neid tuleb hinnata eksperimentaalselt ehk valimi põhjal.

Selle osa eesmärk on ühendada tõenäosusteooria ja matemaatilise statistika vaatenurgad. See tähendab jaotuse parameetrite hindamist Bayesi valemi kontekstis. Kui klassikalisel juhul hindasime uuritava suuruse jaotuse parameetreid valimi põhjal ilma eelneva infota, siis antud juhul lähtume hinnangute juures eelinfost, mida kirjeldab meid huvitava jaotuse parameetri θ eeljaotus. Tähistagem seda kui $p(\theta)$.

Üldistame esmalt Bayesi valemi pidevale juhule. Kui uuritava juhusliku suuruse jaotus on diskreetne, siis

$$P(\theta \in [\theta_1;\theta_2]|X = x_i) = \frac{\int_{\theta_1}^{\theta_2} P(X = x_i|\theta)p(\theta)d\theta}{P(X = x_i)}. \tag{4.5}$$

Pideva jaotuse korral

$$P(\theta \in [\theta_1;\theta_2]|X \leq x) = \frac{\int_{\theta_1}^{\theta_2} f(x|\theta)p(\theta)d\theta}{f(x)}. \tag{4.6}$$

Valemid (4.5) ja (4.6) annavad meile tõenäosuse, et parameeter θ kuulub lõiku $[\theta_1;\theta_2]$ tingimusel, et uuritav suurus $X = x_i$ diskreetsel juhul ning $X \leq x$ pideval juhul. Kui $p(\theta)$ kirjeldab eelteadmist parameetri θ kohta, siis leitud tõenäosus kirjeldab parameetri θ järeljaotust ehk teadmist pärast eksperimenti. Tähistagem järeljaotust kui $\pi(\theta)$. Siis

$$\pi(\theta) = \frac{f(x \mid \theta)p(\theta)}{f(x)}. \tag{4.7}$$

Kuidas aga interpreteerida seoseid (4.5)-(4.6) teaduslikes eksperimentides? Selle interpreteeringu saab kirja panna seosega

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}, \tag{4.8}$$

kus D tähistab andmestikku. Tõenäosus $P(\theta)$ seoses (4.8) iseloomustab parameetri väärtuse θ usalduslikkust ilma andmeteta D. Tõenäosus $P(\theta \mid D)$ aga annab parameetri θ väärtuse usalduslikkuse, kui on arvesse võetud andmestik D. Suurus $P(D \mid \theta)$ tähistab tõepära ehk tõenäosust, et andmestik D esindaks jaotust parameetriga θ. Seda tõepära võib vaadata kui statistilise mudeli tõestust (ingl *evidence of the model*), see tähendab tõestust, et jaotuse parameetri väärtus on θ. Tõenäosust $P(D)$ tuleb käsitleda kui andmete D saamise võimalikkust parameetri θ erinevate väärtuste θ^* korral, see tähendab

$$P(D) = \sum_{\theta^*} P(D \mid \theta^*)P(\theta^*).$$

Enne seoste (4.5) ja (4.6) demonstreerimist erinevate näidetega teeme tutvust kahe jaotusega: beeta- ning gammajaotusega. Tegemist on oluliste jaotustega Bayesi statistikas.

Beetajaotus

Tõenäosusteoorias ja matemaatilises statistikas on beetajaotuseks pidev jaotus, mis on defineeritud lõigul $[0;1]$. Sellele jaotusele vastav tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x \notin (0;1), \\ \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, \text{ kui } x \in (0;1), \end{cases}$$

kus α, $\beta > 0$ ning

$$\mathrm{B}(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

näol on tegemist Euleri beetafunktsiooniga. Beetajaotusele allumist tähistagem kui $X \sim Beta(\alpha,\beta)$. Beetajaotusele vastav keskväärtus ja dispersioon avalduvad järgmiselt:

$$E(X) = \frac{\alpha}{\alpha+\beta} \text{ ning } D(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Beetajaotust kasutatakse paljudes teadusharudes. Seda jaotust on kasutatud kirjeldamaks alleeli sagedust populatsioonigeneetikas. Samuti on beetajaotus rakendust leidnud päikesekiirguse andmete juures, pinnase omaduste uurimisel ning erinevate mineraalide osakaalude uurimisel kivimites.

Gammajaotus

Esitame alljärgnevalt gammajaotuse definitsiooni.

Definitsioon 4.10. Öeldakse, et juhuslik suurus $X \geq 0$ on gammajaotusega, kui tema tihedusfunktsioon

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}\exp(-x\beta)}{\Gamma(\alpha)}, \tag{4.9}$$

kus $\alpha \in \mathcal{R}$ ning $\beta > 0$.

Seega sisaldab gammajaotuse tihedusfunktsioon Euleri gammafunktsiooni. Tähistagem gammajaotusele allumist kui $X \sim G(\alpha, \beta)$. Kui juhuslik suurus X allub gammajaotusele, siis tema keskväärtus ja dispersioon avalduvad järgmiselt:

$$E(X) = \frac{\alpha}{\beta} \text{ ning } D(X) = \frac{\alpha}{\beta^2}.$$

Gammajaotust võib vaadelda, kui eksponentjaotuse üldistust. Seda üldistust saab teha kahel viisil.

1) Kui $\alpha = 1$, siis saame tihedusfunktsioonist (4.9) eksponentjaotuse $\mathcal{E}(\beta)$ tihedusfunktsiooni.

2) Kui juhuslikud suurused $X_1, X_2, ..., X_n$ alluvad eksponentjaotusele $\mathcal{E}(\nu)$, siis juhuslik suurus $Y = \sum_{i=1}^{n} X_i$ allub gammajaotusele $G(n, \nu)$.

Gammajaotus leiab laialdast rakendust kindlustuses kahjunõuete suuruste modelleerimisel. Samuti kasutatakse seda üldistatud lineaarsete mudelite juures vigade hindamisel.

Bayesi valemi rakendusi erinevatele jaotustele

Demonstreerime Bayesi valemit erinevatel jaotustel. Uurime, kuidas mõjutab andmestik D nende jaotuste parameetrite eeljaotusi.

Näide 4.8. Allugu uuritav suurus X Poissoni jaotusele parameetriga M. Hinnatava parameetri M eeljaotuseks olgu eksponentjaotus parameetriga 1. Näites 4.6 saime, et X allub geomeetrilisele jaotusele parameetriga $\frac{1}{2}$. Leiame tingliku tõenäosuse $P(M \leq m \mid X = k)$ ehk uurime, kuivõrd mõjutab parameetri M jaotust teadmine, et juhusliku suuruse X väärtus on k. Tingliku tõenäosuse definitsiooni ja seose (4.6) abil saame, et

$$P(M \leq m \mid X = k) = \frac{P(M \leq m, X = k)}{P(X = k)} =$$

$$= \frac{\int_0^m P(X=k \mid M=y)p(y)dy}{P(X=k)} == \frac{\int_0^m e^{-y}\frac{y^k}{k!}e^{-y}dy}{(\frac{1}{2})^{k+1}} = \int_0^m \frac{y^k 2^{k+1}e^{-2y}dy}{\Gamma(k+1)}.$$

Seega saime tinglikuks jaotusfunktsiooniks

$$F(m \mid x) = \int_0^m \frac{y^k 2^{k+1}e^{-2y}dy}{\Gamma(k+1)}.$$

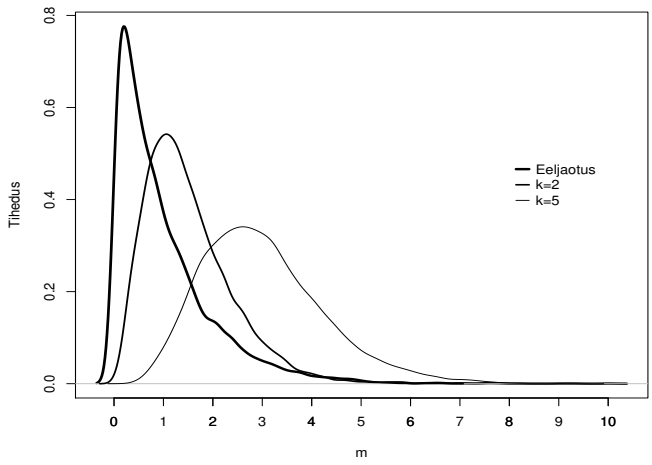
Diferentseerides saame tinglikuks tihedusfunktsiooniks

$$f(m \mid x) = \frac{d}{dm}\int_0^m \frac{y^k 2^{k+1}e^{-2y}dy}{\Gamma(k+1)} = \frac{m^k 2^{k+1}e^{-2m}}{\Gamma(k+1)}, \quad m \geq 0.$$

Kokkuvõttes saime parameetri M järeljaotuseks gammajaotuse $G(k+1, 2)$. Kui on teada, et $X = 5$, siis

$$E(M) = 3 \text{ ja } D(M) = \frac{3}{2}.$$

Jooniselt 4.1 võib visuaalselt jälgida, kuidas mõjutab juhusliku suuruse X väärtuse teadmine parameetri M jaotust. Sellelt jooniselt on näha, et X väärtuse kasvades nihkub parameetri M jaotustiheduse maksimum paremale.



Joonis 4.1. Parameetri M eeljaotus ja järeljaotused $k = 2$ ning $k = 5$ korral

Näite 4.8 eel- ja järeljaotuse mudel on leidnud rakendusi riskiteoorias. Üks nendest rakendustest on toodud ülesandes 4.20.

Näide 4.9. Rakendame valemit (4.6) näitele 4.7. Seal oli binoomjaotuse parameetri N eeljaotuseks Poissoni jaotus parameetriga λ. Seega

$$P(N=n)=\frac{\lambda^n}{n!}e^{-\lambda}.$$

Leiame järeljaotuse

$$P(N=n\mid X=k)=\frac{P(X=k\mid N=n)P(N=n)}{P(X=k)}=$$

$$=\frac{C_n^k p^k q^{n-k}e^{-\lambda}\dfrac{\lambda^n}{n!}}{\dfrac{(\lambda p)^k}{k!}e^{-\lambda p}}=\frac{(\lambda q)^{n-k}e^{-\lambda q}}{(n-k)!},$$

kus $k\le n$. Järeljaotuseks on seega Poissoni jaotus parameetriga λq.

Olgu $X\sim B(N,0.4)$. Eelinfo põhjal olgu teada, et õnnestunud katsete hulk $X=2$. Siis saame kogu katsete hulgale järgmise tingliku tõenäosuse:

$$P(N=n\mid X=2)=\frac{(0.6\lambda)^{n-2}e^{-0.6\lambda}}{(n-2)!}.$$

Antud näite üks rakendusi on toodud ülesandes 4.21.

Näide 4.10. Allugu juhuslik suurus X binoomjaotusele $B(n,Y)$. Katse õnnestumise tõenäosus Y allugu beetajaotusele $Beta(\alpha,\beta)$. Säärane olukord tekib sõltumatute katsete seeria korral, kus katse tulemusel on kaal, mille väärtus kuulub lõiku $[0;1]$. See tähendab, et lisaks õnnestumisele ja ebaõnnestumisele (ehk tulemustele 1 ja 0) on olemas ka vahepealseid variante. Selline olukord esineb näiteks mingi piirkonna metsastuse hindamisel. Hindajal tuleb n proovitüki hulgast leida need, mis kuuluvad metsa alla. Sageli aga on proovitükk mets mingi kaaluga. Sel juhul on üks võimalusi kirjeldada metsaks olemise tõenäosust beetajaotusega.

Rakendades Bayesi valemit, leiame tõenäosuse

$$P(Y\le y\mid X=k)=\frac{P(X=k\mid Y\le y)P(Y\le y)}{P(X=k)}=$$

$$= \frac{\int_0^y t^{k+\alpha-1}(1-t)^{n+\beta-k-1}dt}{\int_0^1 t^{k+\alpha-1}(1-t)^{n+\beta-k-1}dt} = \frac{\int_0^y t^{k+\alpha-1}(1-t)^{n+\beta-k-1}dt}{\mathrm{B}(k+\alpha, n+\beta-k)}.$$

Seega saime seoste (4.6)-(4.7) põhjal parameetri y järeljaotust kirjeldava tihedusfunktsiooni

$$\pi(y) = \frac{y^{k+\alpha-1}(1-y)^{n+\beta-k-1}}{\mathrm{B}(k+\alpha, n+\beta-k)}.$$

Arvutamaks tõenäosust $P(Y \leq y \mid X = k)$ tuleb meil leida integraal

$$I = \int t^{k+\alpha-1}(1-t)^{n+\beta-k-1}dt,$$

mida aga mõningate α ja β väärtuste korral ei eksisteeri analüütilisel kujul. Vaatame ühte lihtsamat juhtu. Olgu parameetrid $\alpha = \beta = 0.5$ ning sõltumatute katsete hulk $n = 5$. Siis saame järgmised tinglikud tõenäosused $P(Y \leq 0.5 \mid X = k)$:

k	0	1	2	3	4	5
$P(Y \leq 0.5 \mid X = k)$	0.993	0.912	0.671	0.332	0.086	0.007

4.2.4. Statistiliste hüpoteeside kontrollimine Bayesi meetodiga

Klassikalise statistika peatükis oli statistiliste hüpoteeside kontrolli peamine eesmärk olulisustõenäosuse (*p-value*) leidmine. Nüüd aga läheneme statistiliste hüpoteeside kontrollile uuest vaatenurgast. Käsitleme hüpoteeside kontrolli täistõenäosuse ja Bayesi valemi kontekstis ning toome sisse uue tõenäosuse.

Tõenäosus *q-value*

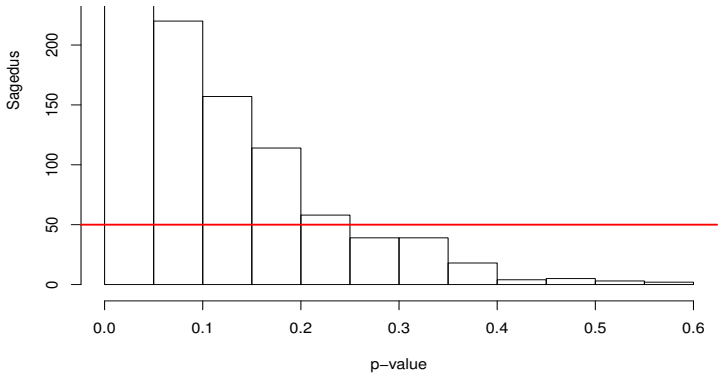
Koosnegu sündmuste süsteem $\mathcal{A}$ nullhüpoteesist H_0 ja sisukast hüpoteesist H_1. Me tahame teadusliku eksperimendiga tõestada sisukat hüpoteesi H_1. Olgu eelneva info põhjal teada, et

$$P(H_0) = \pi_0 \text{ ja } P(H_1) = \pi_1 = 1 - \pi_0.$$

Selle eelteadmisega asume sooritama uusi teaduslikke katseid. Vaatame lähemalt, kuidas leitakse tõenäosust π_0. Olgu meil eelnevalt tehtud m teaduslikku eksperimenti. Igale eksperimendile vastaku üks statistiline test, kus on leitud olulisustõenäosus *p-value*. Eeldame, et kõik testid on sõltumatud. Olulisustõenäosuste jaotuse põhjal koostatakse sageduste histogramm. Sellele tõmmatakse horisontaalne piirjoon kõrgusel $l(\alpha)$. Suurus $l(\alpha)$ on selline statistiliste testide hulk, mille puhul olulisustõenäosus

$$p\text{-}value > \alpha.$$

Tõenäosus α on enamasti väiksem kui 0.5. Üks võimalikke olulisustõenäosuste histogramme on toodud joonisel 4.2.



Joonis 4.2. Olulisustõenäosuste sageduste histogramm koos piirjoonega $m = 1000$, $l(\alpha) = 51$ ning $\alpha = 0.3$ korral

Suuruseks π_0 võetakse kõrgusel $l(\alpha)$ olevast piirjoonest allapoole jäävate tulpade pindalade summa suhe tulpade kogupindalaga.

Põhimõtteliselt võib tõenäosuse π_0 leida ka ilma eelnevate eksperimentideta ehk täielikult teooriale tuginedes.

Toome sisse suuruse, mida tähistame kui PPV (ingl ***P****ositive* ***P****redictive* ***V****alue*). Eesti keeles võiks suurust PPV nimetada positiivseks ennustus-

väärtuseks. Suurus ise avaldub järgmiselt:

$$\mathrm{PPV} = \frac{(1-\pi_0)(1-\gamma_2)}{(1-\pi_0)(1-\gamma_2)+\pi_0\beta}. \tag{4.10}$$

Seega sõltub suurus PPV statistilise testi olulisuse nivoost β ja testi võimsusest $1-\gamma_2$. Lihtne on veenduda, et seos (4.10) sarnaneb Bayesi valemiga. Olgu täistõenäosusele vastav sündmus A positiivse katsetulemuse saamine. Teisisõnu tähendab sündmuse A toimumine nullhüpoteesi kummutamist. Antud juhul

$$\beta = P(A \mid H_0) \text{ ja } 1-\gamma_2 = P(A \mid H_1).$$

Suurus

$$\mathrm{PPV} = P(H_1 \mid A),$$

kus täistõenäosus

$$P(A) = (1-\pi_0)(1-\gamma_2)+\pi_0\beta.$$

Koosnegu meie teadusekseperiment m katsest. Olgu m_1 katsete hulk, mille puhul toimus sündmus A ning m_0 katsete hulk, mille puhul toimus selle sündmuse vastandsündmus. Seega $m = m_0 + m_1$. Suurus PPV näitab tõeste H_1 efektide osakaalu m_1 katse hulgas. Seda osakaalu võib tõlgendada kui tõenäosust, et saadud positiivne katsetulemus on ka õige. Lahtiseletatult – me saime positiivse katsetulemuse ehk teame, et sündmus A on toimunud. Kui suure tõenäosusega on siis õige hüpotees H_1?

Järgnevalt uurime, kuidas leida valede positiivsete tulemuste tõenäosust. Selleks tuuakse sisse suurus, mida tähistatakse kui FDR (ingl ***F**alse **D**iscovery **R**ate*). Selle suuruse eestikeelne nimetus võiks olla vale avastuse määr. Suurus avaldub järgmiselt:

$$\mathrm{FDR} = \frac{\pi_0\beta}{\pi_0\beta+(1-\pi_0)(1-\gamma_2)}. \tag{4.11}$$

Seosest (4.11) ning Bayesi valemist järeldub, et

$$\mathrm{FDR} = P(H_0 \mid A).$$

Seega iseloomustab suurus FDR valede H_1 efektide osakaalu m_1 katse hulgas. Seoste (4.10) ja (4.11) põhjal saame, et

$$\mathrm{PPV} + \mathrm{FDR} = P(H_1 \mid A) + P(H_0 \mid A) = 1.$$

Näide 4.11. Enne dopingukontrolli oli põhjust eeldada, et 10 % sportlastest tarvitab keelatuid aineid. Dopingutarvitajate avastamiseks viidi läbi test, mille võimsuseks oli keskmiselt 0.95. Kuna selle testi tulemusest sõltus sportlase saatus, siis võeti olulisuse nivooks $\beta = 0.01$. Kui suure tõenäosusega oli positiivse testi tulemuse saanud sportlane dopingupatune?

Olgu sündmus $A =$ „testi tulemus on positiivne“. Antud juhul $\pi_0 = 0.9$ ja $\pi_1 = 0.1$. Seose (4.10) põhjal saame, et

$$\mathrm{PPV} = P(H_1 \mid A) = \frac{0.95 \cdot 0.1}{0.95 \cdot 0.1 + 0.01 \cdot 0.9} \approx 0.913.$$

Seega vale avastuse määr $\mathrm{FDR} = 0.087$. Seega võime väita, et selle testi põhjal ei saa otsustada sportlase saatuse üle. Vale positiivsuse risk on selleks liiga suur.

Näide 4.12. Uurime suurusi PPV ning FDR õigusteaduslikul ehk formaaljuriidilisel tasandil. Kuriteos kahtlustatava isiku puhul püstitatakse hüpoteeside paar

$$\begin{cases} H_0 : \text{isik on süütu}, \\ H_1 : \text{isik on süüdi}. \end{cases}$$

Eelinfo põhjal on süüdi olemise tõenäosus π_1. Kontrollimaks süüdistust saadakse tõendusmaterjal, mida võib vaadelda kui sündmusena A. See sündmus võib aga toimuda nii süüdi oleva kui ka süütu inimese puhul ehk

$$P(A) = P(A \mid H_1)\pi_1 + P(A \mid H_0)(1 - \pi_1).$$

Peale tõendusmaterjali esitamist on isiku süüdi olemise tõenäosus

$$\mathrm{PPV} = P(H_1 \mid A) = \frac{P(A \mid H_1)\pi_1}{P(A)}.$$

Leitud tõenäosus PPV on uurimisorganite töö tulemus, mille nad esitavad kohtuorganitele. Kohus aga peab olema viimse hetkeni kinni nullhüpoteesi H_0 juures ning langetama otsuse olulisustõenäosuse *p-value* põhjal.

Suurusel FDR põhineb tõenäosus, mille ingliskeelseks nimetuseks on *q-value*. Selle suuruse näol on tegemist olulisustõenäosuse (*p-value*) n-ö

Bayesi versiooniga. Tähistagu D saadud andmeid. Siis

$$p\text{-}value = P(D \mid H_0) \text{ ning } q\text{-}value = P(H_0 \mid D).$$

Anname järgnevalt suurusele *q-value* matemaatilise määratluse. Seda saab teha tuginedes John D. Storey teadusartiklile [39]. Esmalt defineeritakse piirkondade süsteem $\{\mathcal{H}_\alpha\}_{\alpha=0}^{1}$, kus α on selline tõenäosus, et

$$P(T(\mathbf{x}) \in \mathcal{H}_\alpha \mid H_0) = \alpha. \tag{4.12}$$

Kui $\alpha' \leq \alpha$, siis $\mathcal{H}_{\alpha'} \subseteq \mathcal{H}_\alpha$. Olulisustõenäosuse (*p-value*) puhul on tegemist valemi (4.12) erijuhuga. Igale piirkonnale $\mathcal{H}_\alpha$ saab seada vastavusse mingi vale avastuse määra

$$\mathrm{FDR}(\mathcal{H}_\alpha) = P(H_0 \mid T(\mathbf{x}) \in \mathcal{H}_\alpha).$$

Nende leitud $\mathrm{FDR}(\mathcal{H}_\alpha)$ väärtuste põhjal defineeritakse tõenäosus *q-value*.

Definitsioon 4.11. Teststatistiku $T(\mathbf{X})$ väärtusele t vastav tõenäosus

$$q\text{-}value(t) = \min_{\{\mathcal{H}_\alpha : t \in \mathcal{H}_\alpha\}} \mathrm{FDR}(\mathcal{H}_\alpha).$$

Seega on tõenäosuse *q-value* arvutamise puhul tegemist lokaalse miinimumi leidmise probleemiga. Tuleb leida selline teststatistiku väärtus t, mille puhul $\mathrm{FDR}(\mathcal{H}_\alpha) \to \min$. Võib öelda, et tõenäosus *q-value* mõõdab teststatistiku $T(\mathbf{X})$ kindlust vale avastuse tegemise vastu.

Leidmaks tõenäosust *q-value* tarkvara R abil tuleb esmalt anda järgmised käsud:

```
source("https://bioconductor.org/biocLite.R")
biocLite("qvalue").
```

Seejärel tuleb laadida R-i tööpakett `qvalue`.

Näide 4.13. Olgu eelnevate testide olulisustõenäosused järgmised: 0.01, 0.05, 0.15 ja 0.9. Siis saame leida suuruse q-value väärtused käskudega

```
p=c(0.01,0.05,0.15,0.9)
qvalue(p).
```

Väljundiks saame tabeli

p-value	0.01	0.05	0.15	0.9
q-value	0.025	0.063	0.13	0.57

Antud juhul on eelnevate olulisustõenäosuste põhjal saadud $\pi_0 = 0.635$.

Statistiliste hüpoteeside kontrolli puhul saab formuleerida kaks erinevat strateegiat.

1) Olulisustõenäosuse (*p-value*) leidmise strateegia. See strateegia põhineb eeldusel, et nullhüpotees H_0 on õige. Leitakse, kui suur on kogutud andmete saamise tõenäosus nullhüpoteesi eelduse korral.

2) Eelinfo kasutamise strateegia. Enne eksperimenti on meil olemas varasema info põhjal saadud eeltõenäosused (ehk eeljaotus) $\pi_0 = P(H_0)$ ja $\pi_1 = P(H_1)$ näol. Leiame, kuivõrd mõjutab saadud andmestik neid tõenäosusi. Otsustus tehakse erinevate FDR-i näitajate baasil leitud *q-value* põhjal.

Esimene strateegia eeldab kvaliteetseid andmeid, teine aga adekvaatset eelinfot H_0 ja H_1 kohta. Kui tegemist on vastutusrikka testiga, siis tuleks eelistada esimest strateegiat. Teine strateegia annab aga rohkem infot, sest tõenäosus *q-value* sisaldab endas nii teooria kui ka eksperimendi tulemusi.

Tõepära suhe

Uurime statistiliste hüpoteeside kontrolli meetodit, milles on ühendatud *p-value* ja *q-value* leidmise strateegia. Selleks defineerime esmalt suuruse, mida nimetatakse tõepära suhteks. Edaspidi kasutame selle suuruse ingliskeelset lühendit LR (***L****ikelihood* ***R****atio*). Eelnevalt toodud tähistustes saame tõepära suhte defineerida järgmiselt:

$$\mathrm{LR} = \frac{P(H_1 \mid D)P(H_0)}{P(H_0 \mid D)P(H_1)} = \frac{P(D \mid H_1)}{P(D \mid H_0)}.$$

See suhe on üks kriteeriumitest langetamaks valikut kas H_0 või H_1 kasuks. Mida suurem on suhte LR väärtus, seda rohkem on põhjust langetada otsus hüpoteesi H_1 kasuks. Demonstreerime tõepära suhte leidmist lihtsa näite abil.

Näide 4.14. Olgu kastis 8 detaili. Nende detailide kohta esitatakse kaks hüpoteesi:

1) $H_0 =$ „detailide seas on 2 praakdetaili“,

2) $H_1 =$ „detailide seas on 3 praakdetaili“.

Kastist võeti juhuslikult 3 detaili. Juhusliku katse tulemuseks saadi sündmus

$$D = \text{„testitud detailidest osutus 1 praagiks ja 2 korras olevaks“}.$$

Kuivõrd mõjutab katse tulemus D sündmuste H_0 ja H_1 tõenäosusi? Leiame tõepära suhte

$$\mathrm{LR} = \frac{P(D \mid H_1)}{P(D \mid H_0)} = \frac{\dfrac{C_3^1 C_5^2}{C_8^3}}{\dfrac{C_2^1 C_6^2}{C_8^3}} = \frac{3 \cdot 10}{2 \cdot 15} = 1.$$

Seega ei kalluta katse tulemus D eelistust ei H_0 ega H_1 suunas.

Käsitleme nüüd tõepära suhet statistiliste hüpoteeside kontekstis. Selleks defineerime suuruse LR kui tõepära funktsioonide suhte

$$\mathrm{LR} = \frac{L(\mathbf{x} \mid H_1)}{L(\mathbf{x} \mid H_0)}.$$

Antud juhul on tõepära funktsiooni näol tegemist tinglike tiheduste korrutisega

$$L(\mathbf{x} \mid H_1) = \prod_{i=1}^{n} f(x_i \mid H_1) \text{ ning } L(\mathbf{x} \mid H_0) = \prod_{i=1}^{n} f(x_i \mid H_0).$$

Rakendame statistilise hüpoteeside kontrolli juures Neyman-Pearsoni lemmat. Sõnastame selle lemma teoreemina.

Teoreem 4.3. (Neymann-Pearsoni lemma) Olgu meil hüpoteeside paar

$$\begin{cases} H_0 : \theta = \theta_0, \\ H_1 : \theta = \theta_1. \end{cases}$$

Siis tõepära suhte test, mis kummutab hüpoteesi H_0 tingimusel

$$\mathrm{LR} = \frac{L(\mathbf{x} \mid \theta_1)}{L(\mathbf{x} \mid \theta_0)} \geq l,$$

kus

$$P(\mathrm{LR} \geq l \mid H_0) = \beta$$

on võimsaim statistiline test olulisuse nivool β.

Tõestus. Defineerime esmalt hüpoteesi H_0 kummutamise piirkonna $\mathcal{H}_{NP} = \left\{\mathbf{x} : \frac{L(\mathbf{x} \mid \theta_1)}{L(\mathbf{x} \mid \theta_0)} \geq l\right\}$, kus suurus l on valitud selliselt, et $P(\mathcal{H}_{NP} \mid \theta_0) = \beta$. Olgu $\mathcal{H}_*$ muu suvaline H_0 kummutamise piirkond. Vastavalt otsustuse kriteeriumile peab piirkond $\mathcal{H}_*$ olema selline, et $P(\mathcal{H}_* \mid \theta_0) \leq \beta$. Seega

$$\beta = P(\mathcal{H}_{NP} \mid \theta_0) \geq P(\mathcal{H}_* \mid \theta_0).$$

Olgu $\mathcal{H}_{NP}^c$ ning $\mathcal{H}_*^c$ piirkondade $\mathcal{H}_{NP}$ ja $\mathcal{H}_*$ täiendpiirkonnad. Neis tähistustes saame, et

$$P(\mathcal{H}_{NP} \mid \theta) = P(\mathcal{H}_{NP} \cap \mathcal{H}_* \mid \theta) + P(\mathcal{H}_{NP} \cap \mathcal{H}_*^c \mid \theta)$$

ning

$$P(\mathcal{H}_* \mid \theta) = P(\mathcal{H}_{NP} \cap \mathcal{H}_* \mid \theta) + P(\mathcal{H}_{NP}^c \cap \mathcal{H}_* \mid \theta).$$

Võttes $\theta = \theta_0$, saame nende kahe võrduse ja ülal toodud võrratuse põhjal, et

$$P(\mathcal{H}_{NP} \cap \mathcal{H}_*^c \mid \theta_0) \geq P(\mathcal{H}_{NP}^c \cap \mathcal{H}_* \mid \theta_0).$$

Võrdleme 2 testi võimsust. Vastavalt testi võimsuse definitsioonile tuleb võrrelda tõenäosusi $P(\mathcal{H}_{NP} \mid \theta_1)$ ning $P(\mathcal{H}_* \mid \theta_1)$. Teoreemis esitatud väite tõestuseks peab näitama, et

$$P(\mathcal{H}_{NP} \mid \theta_1) \geq P(\mathcal{H}_* \mid \theta_1) \Longleftrightarrow P(\mathcal{H}_{NP} \cap \mathcal{H}_*^c \mid \theta_1) \geq P(\mathcal{H}_{NP}^c \cap \mathcal{H}_* \mid \theta_1).$$

Arvestades eeldust, et $L(\mathbf{x} \mid \theta_1) \geq lL(\mathbf{x} \mid \theta_0)$, saame, et

$$P(\mathcal{H}_{NP} \cap \mathcal{H}_*^c \mid \theta_1) = \int_{\mathcal{H}_{NP} \cap \mathcal{H}_*^c} L(\mathbf{x} \mid \theta_1) d\mathbf{x} \geq l \int_{\mathcal{H}_{NP} \cap \mathcal{H}_*^c} L(\mathbf{x} \mid \theta_0) d\mathbf{x} =$$

$$= lP(\mathcal{H}_{NP} \cap \mathcal{H}_*^c \mid \theta_0) \geq lP(\mathcal{H}_{NP}^c \cap \mathcal{H}_* \mid \theta_0) = l \int_{\mathcal{H}_{NP}^c \cap \mathcal{H}_*} L(\mathbf{x} \mid \theta_0) d\mathbf{x} \geq$$

$$\geq \int_{\mathcal{H}_{NP}^c \cap \mathcal{H}_*} L(\mathbf{x} \mid \theta_1) d\mathbf{x} = P(\mathcal{H}_{NP}^c \cap \mathcal{H}_* \mid \theta_1).$$

□

Nullhüpoteesi kummutamise kriteeriumit LR $\geq l$ nimetatakse Neymann-Pearsoni kriteeriumiks.

Uurime, milline on parim lävendi väärtus l. Defineerime järgmised sündmused:

1) sündmus A – tõene on hüpotees H_0;

2) sündmus $\overline{A}$ – tõene on hüpotees H_1;

3) sündmus B – katsetulemus x sattus nullhüpoteesi piirkonda $\mathcal{H}_0$;

4) sündmus $\overline{B}$ – katsetulemus x sattus sisuka hüpoteesi piirkonda $\mathcal{H}_1$.

Katsetulemuse x näol on tegemist juhusliku suuruse X mingi realisatsiooniga. Olgu juhuslikul suurusel X kaks teadaolevat tinglikku tihedusfunktsiooni:

1) $f_0(x)$ – tingimusel, et tõene on hüpotees H_0;

2) $f_1(x)$ – tingimusel, et tõene on hüpotees H_1.

Sündmuste korrutistele AB ja $\overline{A}\overline{B}$ vastab õige otsustus, korrutisele $A\overline{B}$ vastab I liiki viga ning korrutis $\overline{A}B$ on seotud II liiki veaga. Defineerime juhusliku suuruse C, mis iseloomustab valest otsusest tingitud kahju mingis ühikus. Olgu

$$C = \begin{cases} 0, & \text{kui toimus } AB \text{ või } \overline{A}\overline{B}, \\ c_1, & \text{kui toimus } A\overline{B}, \\ c_2, & \text{kui toimus } \overline{A}B. \end{cases}$$

Juhusliku suuruse C jaotus olgu järgmine: $P(C = 0) = p_0$, $P(C = c_1) = p_1$ ning $P(C = c_2) = p_2$. Tähistagem keskmist kahjumit kui r. Seega

$$r = E(C) = 0p_0 + c_1p_1 + c_2p_2 = c_1p_1 + c_2p_2.$$

Eesmärk on leida optimaalne otsustuse kriteerium, mille puhul keskmine kahjum r oleks minimaalne. Avaldame I ja II liiki vea tegemise tõenäosuses tinglike jaotustiheduste $f_0(x)$ ning $f_1(x)$ kaudu

$$\gamma_1 = P(\overline{B} \mid A) = \int_{\mathcal{H}_1} f_0(x)dx \text{ ja } \gamma_2 = P(B \mid \overline{A}) = \int_{\mathcal{H}_0} f_1(x)dx.$$

Järgnevalt avaldame tõenäosused p_1 ja p_2 tõenäosuste γ_1 ja γ_2 ning aprioorse tõenäosuse π_0 kaudu. Saame

$$p_1 = P(\overline{B}A) = P(A)P(\overline{B} \mid A) = \pi_0\gamma_1$$

ning

$$p_2 = P(\overline{A}B) = P(\overline{A})P(B \mid \overline{A}) = (1-\pi_0)\gamma_2.$$

Kui asendame saadud tõenäosused p_1 ning p_2 keskmise kao valemisse, siis saame

$$r = \pi_0\gamma_1 c_1 + (1-\pi_0)\gamma_2 c_2 = \pi_0 c_1 \int_{\mathcal{H}_1} f_0(x)dx + (1-\pi_0)c_2 \int_{\mathcal{H}_0} f_1(x)dx.$$

Seega tuleb meil leida piirkonna $\mathcal{H}$ selline tükeldus piirkondadeks $\mathcal{H}_0$ ja $\mathcal{H}_1$, et keskmine kahjum r oleks minimaalne. Arvestades jaotustiheduse aditiivsuse omadust

$$\int_{\mathcal{H}} f_0(x)dx = \int_{\mathcal{H}_0} f_0(x)dx + \int_{\mathcal{H}_1} f_0(x)dx = 1,$$

saame

$$r = \pi_0 c_1 \left(1 - \int_{\mathcal{H}_0} f_0(x)dx\right) + (1-\pi_0)c_2 \int_{\mathcal{H}_0} f_1(x)dx =$$

$$= \pi_0 c_1 + \int_{\mathcal{H}_0} \{(1-\pi_0)c_2 f_1(x) - \pi_0 c_1 f_0(x)\}dx.$$

Saadud avaldise väärtus on minimaalne sellises nullhüpoteesi $\mathcal{H}_0$ piirkonnas, mis minimeerib integraali võrduse parema poole. Selleks aga tuleb piirkonda $\mathcal{H}_0$ lugeda need ja ainult need x väärtused, mille puhul määratud integraal on negatiivne. Selle negatiivsus on garanteeritud, kui

$$(1-\pi_0)c_2 f_1(x) - \pi_0 c_1 f_0(x) < 0$$

ehk

$$\frac{f_1(x)}{f_0(x)} < \frac{\pi_0 c_1}{(1-\pi_0)c_2}. \tag{4.13}$$

Suhet $\dfrac{f_1(x)}{f_0(x)}$ võrratuses (4.13) nimetatakse tõepära suhteks. Tähistades seda suhet kui l, saame järgmise otsustuse kriteeriumi:

$$l = \frac{\pi_0 c_1}{(1-\pi_0)c_2}.$$

Kui tõepära suhe on väiksem kui suurus l, siis oleme sunnitud jääma nullhüpoteesi H_0 juurde. Vastasel juhul aga loeme tõestatuks sisuka hüpoteesi H_1.

Rakendame saadud tulemuse binoomjaotusele.

Näide 4.15. Näide on seotud toodangu partii kvaliteediga. Küsimus on selles, kas lasta partii müüki või mitte. Selleks püstitati järgmine hüpoteeside paar:

$$\begin{cases} H_0 : \text{partii ei ole müügikõlblik}, \\ H_1 : \text{partii on müügikõlblik}. \end{cases}$$

Hüpoteesi H_0 korral olgu suurem praagi tõenäosus $p = p_0$. Hüpoteesi H_1 puhul siis väiksem praagi risk $p = p_1$.

Kontrolliti sõltumatult n toodet. Olgu x praaktoodete hulk, mis on juhusliku suuruse $X \sim B(n,p)$ mingi realisatsioon. Vastavalt binoomjaotuse jaotusseadusele saadi järgmine tõepära suhe:

$$\frac{f_1(x)}{f_0(x)} = \frac{P(X = x \mid p = p_1)}{P(X = x \mid p = p_0)} = \frac{C_n^x p_1^x (1-p_1)^{n-x}}{C_n^x p_0^x (1-p_0)^{n-x}} = \\ = \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^n,$$

Võrratuse (4.13) põhjal saame, et

$$\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^n < \frac{\pi_0 c_1}{(1-\pi_0)c_2},$$

millest

$$\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x < \frac{\pi_0 c_1}{(1-\pi_0)c_2} \left(\frac{1-p_0}{1-p_1}\right)^n. \tag{4.14}$$

Eeldusel, et nullhüpoteesi korral on praagi tegemise risk suurem, lisati tingimus $p_0 > p_1$. Seega

$$\frac{p_1(1-p_0)}{p_0(1-p_1)} < 1 \Leftrightarrow \ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) < 0.$$

Võrratuse (4.14) põhjal saame, et

$$x \ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) < \ln\left(\frac{\pi_0 c_1}{(1-\pi_0)c_2}\left(\frac{1-p_0}{1-p_1}\right)^n\right) \Leftrightarrow$$

$$\Leftrightarrow x > \frac{\ln\left(\frac{\pi_0 c_1}{(1-\pi_0)c_2}\left(\frac{1-p_0}{1-p_1}\right)^n\right)}{\ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)} := l.$$

Seega, kui n toote seast osutus praagiks rohkem kui l toodet, siis võetakse vastu otsus, et partii on halva kvaliteediga ning müügikõlbmatu. Vastasel juhul aga lastakse see müüki.

Olgu sisuka hüpoteesi H_1 korral praagi tõenäosus $p_1 = 0.05$. Nullhüpoteesi H_0 puhul aga olgu $p_0 = 0.5$. Eelnevate katsete põhjal on leitud, et 80%-l juhtudest tuleb jääda nullhüpoteesi juurde. Seega aprioorne tõenäosus $\pi_0 = 0.8$. Eeldati, et I liiki vea tegemine on II liiki vea tegemisest 3 korda kallim, seega $\frac{c_1}{c_2} = 3$. See tähendab, et trahv praaktoote müügile laskmise eest on 3 korda suurem, kui korras toodete müümata jäämisest saamata jäänud kasum. Kvaliteedikontroll viidi sõltumatult läbi 100 toote juures. Selliste tingimuste juures tehakse järgmine otsustus: kui üle 9 toote osutus praagiks, siis ei ole partii müügikõlblik. Vastasel juhul aga võib lasta partii müügile.

Ülesanne lugejatele: leidke antud näitele vastava suuruse l väärtus.

Rakendame tõepära suhte kriteeriumit normaaljaotusele.

Näide 4.16. Antud näites testiti mingi suuruse Y erinevust keskmisest. Näiteks

1) kuu keskmise temperatuuri erinevust paljude aastate keskmisest;

2) mingi keemilise ühendi reostuskoormuse kõrvalekallet keskmisest;

3) detaili tegeliku läbimõõdu erinevust õigest läbimõõdust.

Olgu suuruse Y keskmine (või õige) näitaja μ. Tavapärasest kõrvale kaldunud näitaja $Y = \mu + X$. Olgu $X \sim \mathcal{N}(0, \sigma)$, siis $Y \sim \mathcal{N}(\mu, \sigma)$. Püstitati hüpoteesid

$$\begin{cases} H_0 : E(Y) - \mu = 0, \\ H_1 : E(Y) - \mu \neq 0. \end{cases}$$

Tõepära suhe

$$\frac{f_1(y)}{f_0(y)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(y-\mu-x)^2}{2\sigma^2} \right)}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(y-\mu)^2}{2\sigma^2} \right)} = \exp\left(-\frac{x^2}{2\sigma^2} \right) \exp\left(\frac{x(y-\mu)}{\sigma^2} \right).$$

Võrratusest (4.13) saame, et

$$\exp\left(- \frac{x^2}{2\sigma^2} \right) \exp\left(\frac{x(y-\mu)}{\sigma^2} \right) < \frac{\pi_0 c_1}{(1-\pi_0)c_2}.$$

Pärast mõningaid teisendusi saame võrratuse

$$\mid y-\mu \mid < \frac{\sigma^2}{\mid x \mid} \left| \ln\left(\frac{\pi_0 c_1}{(1-\pi_0)c_2} \exp\left(\frac{x^2}{2\sigma^2} \right) \right) \right| := l. \tag{4.15}$$

Kui näitaja Y väärtuseks saadi selline y, mis rahuldab võrratust (4.15), siis tuleb jääda nullhüpoteesi H_0 juurde. Vastasel korral lugeda tõestatuks H_1.

Vaatame järgmist hüpoteeside paari:

$$\begin{cases} H_0 : \text{mõõtmine on ilma süstemaatilise veata,} \\ H_1 : \text{mõõtmises esineb süstemaatiline viga.} \end{cases}$$

Olgu mõõdetava suuruse y õige väärtus μ. Süstemaatilise vea ehk nihke korral oleks $y = \mu + x$. Seega tuleb testida, kas suurus x on 0 või sellest erinev. Olgu juhusliku suuruse Y standardhälve $\sigma = 0.2$ ühikut.Tehtud 8 mõõtmise põhjal saadi nihke aritmeetiliseks keskmiseks $\overline{x} = 0.1125$, mis on juhusliku suuruse $\overline{X} \sim \mathcal{N}\left(0, \frac{0.2}{\sqrt{8}}\right)$ mingi realisatsioon. Eelnevate mõõtmiste põhjal on teada, et nihe esineb 10% juhtudel. Eeldades, et kahjud on I ja II liiki vea puhul võrdsed ($c_1 = c_2$), saame

$$l = \frac{0.2^2}{8 \cdot 0.1125} \ln\left(\frac{0.9}{1-0.9} \right) \exp\left(\frac{8 \cdot 0.1125^2}{2 \cdot 0.2^2} \right) \approx 0.0931.$$

Seega

$$\mid y-\mu \mid = 0.1125 > l$$

ning me saame tõestada hüpoteesi H_1. Järelikult esines mõõtmises oluline süstemaatiline viga.

4.3. Markovi ahelad statistikas

Käsitleme meetodit nimega MCMC (ingl ***M**arkov **C**hain **M**onte **C**arlo*). Tegemist on jaotuse parameetrite hinnangu parandamisega eel- ja järeljaotuse abil. See hinnang sisaldab endas nii teoreetilisi eelteadmisi kui ka empiirikat. Selliseks hinnanguks rakendadakse Markovi ahelaid. Esmalt aga tutvume Bayesi statistika põhiteoreemiga.

4.3.1. Bayesi teoreem

Järgnevalt teeme tutvust teoreemiga, millel põhineb Bayesi statistika. Bayesi statistika koosneb eelinfost ja mõõtmistulemustest. Eelinfoks on mingi jaotuse parameetrite vektori $\Theta = (\theta_1, \theta_2, ..., \theta_k)^\intercal$ eeljaotus $p(\Theta)$. Tuleme tagasi seostega (1.5) ja (1.6) defineeritud tõepära funktsiooni L juurde. Tõepära funktsiooni abil leitakse parameetrite vektori Θ parim hinnang valimi $\mathbf{X}$ mingi realisatsiooni korral. Seega sisaldab tõepära funktsioon mõõtmistulemuste infot. Funktsiooni L saab avaldada tingliku tiheduse kaudu järgmiselt:

$$L(\Theta) = f(\mathbf{x}|\Theta) = \prod_{i=1}^{n} f(x_i|\Theta),$$

kus $\mathbf{x}$ tähistab valimi mingit realisatsiooni. Osutub, et tõepärafunktsioon mängib tähtsat rolli Bayesi statistikas. Sissejuhatuseks vaatame tuntud näidet.

Näide 4.17. Käesolev näide puudutab juhuviga. Olgu θ meid huvitav konstant ning ϵ_i, $i = 1, 2, ..., n$, iseloomustagu juhuviga. Eeldame, et $\epsilon_i \sim \mathcal{N}(0, \sigma)$. Siis avalduvad vaatlused kui juhuslikud suurused $X_i = \theta + \epsilon_i$. Kui need juhuslikud suurused on sõltumatud, siis $X_i \sim \mathcal{N}(\theta, \sigma)$. Antud juhul $\Theta = (\theta, \sigma)^\intercal$ ning tõepärafunktsioon

$$L(\Theta) = f(\mathbf{x}|\Theta) = \prod_{i=1}^{n} f(x_i|\theta, \sigma) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x_i - \theta)^2}{2\sigma^2} \right) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right).$$

Näite 4.17 puhul saame leida 0.95-usaldusintervalli $I_{0.95}$ parameetrile θ klassikalisel moel

$$I_{0.95} = \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

Selles usaldusintervallis on arvesse võetud üksnes juhuviga. Süstemaatilise vea puhul see vahemikhinnang ei toimi. Sel juhul tuleb rakendada tulemust, mida nimetatakse Bayesi teoreemiks.

Enne Bayesi teoreemi juurde asumist toome sisse proportsionaalsuse tähistuse. Olgu c argumendist x sõltumatu konstant. Siis tähistame seose $\phi_1(x) = c\phi_2(x)$ kui $\phi_1(x) \propto \phi_2(x)$. Seose $\propto$ puhul öeldakse, et funktsioonid ϕ_1 ja ϕ_2 on proportsionaalsed.

Teoreem 4.4. (Bayesi teoreem) Olgu parameetrite vektori Θ järeljaotus

$$\pi(\Theta) = p(\Theta \mid \mathbf{x}) = \frac{f(\mathbf{x}, \Theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x} \mid \Theta)p(\Theta)}{f(\mathbf{x})},$$

kus

$$f(\mathbf{x}) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}}_{k \text{ korda}} f(\mathbf{x} \mid \Theta)p(\Theta)d\theta_1 d\theta_2 ... d\theta_k.$$

Siis $\pi(\Theta) \propto L(\Theta)p(\Theta)$.

Tõestus. Olgu meil esimese eksperimendi tulemused $\mathbf{x}_1$, funktsioon $L_1(\Theta)$ olgu nendele tulemustele vastav tõepärafunktsioon ning $\pi_1(\Theta)$ olgu pärast neid vaatlusi saadud järeljaotus. Olgu teise eksperimendi vaatlustulemused $\mathbf{x}_2$. Nende vaatluste jaotust iseloomustagu tõepärafunktsioon $L_2(\Theta)$. Peale vaatlustulemusi $\mathbf{x}_2$ saadud järeljaotus

$$\pi_2(\Theta) = p(\Theta|\mathbf{x}_2) = \frac{f(\mathbf{x}_2 \mid \Theta)\pi_1(\Theta \mid \mathbf{x}_1)}{f(\mathbf{x}_2)} = \frac{f(\mathbf{x}_2 \mid \Theta)f(\mathbf{x}_1 \mid \Theta)p(\Theta)}{f(\mathbf{x}_2)f(\mathbf{x}_1)}.$$

Seega saime, et

$$\pi_2(\Theta) \propto L_2(\Theta)L_1(\Theta)p(\Theta).$$

Jätkates antud eksperimenti m korda, saame, et

$$\pi_m(\Theta) \propto L(\Theta)p(\Theta),$$

kus

$$L(\Theta) = L_m(\Theta)L_{m-1}(\Theta)...L_2(\Theta)L_1(\Theta) = \prod_{j=1}^{m} L_j(\Theta).$$

Seega $\pi(\Theta) \propto L(\Theta)p(\Theta)$.

□

Eeljaotuse $p(\Theta)$ ning järeljaotuste $\pi_i(\Theta)$, $i = 1, 2, ..., m$, parameetreid nimetatakse hüperparameetriteks. Üldistame Bayesi teoreemi valguses näidet 4.17 juhule, mil peale juhuvea esineb ka süstemaatiline viga.

Näide 4.18. Olgu meil $X_i \sim \mathcal{N}(\theta, \sigma)$ ning meid huvitav suurus $\theta \sim \mathcal{N}(\mu, \tau)$. Antud juhul on tegemist nullitriiviga ehk olukorraga, kus mõõteriista skaala näitu mõjutavad segavad tegurid (näiteks välistemperatuuri muutused, õhuniiskus jms). Juhuslik suurus

$$X_i = \theta + \epsilon_{i1} + \epsilon_{i2},$$

kus $\epsilon_{i1} \sim \mathcal{N}(0, \sigma)$ iseloomustab juhuviga ning $\epsilon_{i2} \sim \mathcal{N}(\mu - \theta, \tau)$ süstemaatilist viga. Juhuslikud suurused ϵ_{i1} ja ϵ_{i2} ei pruugi olla sõltumatud. Meie eesmärk on leida suuruse θ jaotus pärast n mõõtmist. Eeljaotus

$$p(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(- \frac{(\theta - \mu)^2}{2\tau^2} \right).$$

Tõepärafunktsioon

$$L(\theta) \propto \prod_{i=1}^{n} \exp\left(- \frac{(x_i - \theta)^2}{2\sigma^2} \right) = \exp\left(- \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} \right) =$$

$$= \exp\left(- \sum_{i=1}^{n} \frac{(x_i - \overline{x} + \overline{x} - \theta)^2}{2\sigma^2} \right) =$$

$$= \exp\left(- \frac{\sum_{i=1}^{n} \left\{ (x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - \theta) + (\overline{x} - \theta)^2 \right\}}{2\sigma^2} \right) =$$

$$= \exp\left(- \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{2\sigma^2} \right) \exp\left(- n\frac{(\overline{x} - \theta)^2}{2\sigma^2} \right) \propto \exp\left(- n\frac{(\overline{x} - \theta)^2}{2\sigma^2} \right).$$

Bayesi teoreemi põhjal saame, et järeljaotus

$$\pi(\theta) \propto L(\theta)p(\theta) \propto \prod_{i=1}^{n} f(x_i \mid \theta)p(\theta) \propto$$

$$\propto \exp\left(-n\frac{(\overline{x}-\theta)^2}{2\sigma^2}\right)\exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right) =$$

$$= \exp\left(-\frac{1}{2}\left\{ \frac{n(\overline{x}-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu)^2}{2\tau^2}\right\}\right) =$$

$$= \exp\left(-\frac{1}{2}\left\{ \frac{n\overline{x}^2}{\sigma^2} - \frac{2n\overline{x}\theta}{\sigma^2} + \frac{n\theta^2}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\theta\mu}{\tau^2} + \frac{\mu^2}{\tau^2}\right\}\right).$$

Leidmaks mõõtmistulemuste mõju järeljaotusele tuuakse sisse suurused

$$\tau_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \text{ ja } \mu_1 = \tau_1^2\left(\frac{n\overline{x}}{\sigma^2} + \frac{\mu}{\tau^2}\right).$$

Nende suuruste põhjal saame, et

$$\pi(\theta) \propto \exp\left(-\frac{1}{2}\left\{ \frac{n(\overline{x}-\mu)^2}{\sigma^2+n\tau^2} + \frac{(\theta-\mu_1)^2}{\tau_1^2}\right\}\right) \propto \exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right).$$

Seega on järeljaotus $\pi(\theta)$ proportsionaalne normaaljaotusega $\mathcal{N}(\mu_1, \tau_1)$. Selle järeljaotuse põhjal saab konstrueerida α-usaldusintervalli parameetrile θ.

Vaatame näidet eksponentjaotuse parameetri eel- ja järeljaotusest.

Näide 4.19. Olgu $X_i \sim \mathcal{E}(\nu)$, $i = 1, 2, ..., n$. Olgu parameetri ν eeljaotuseks ühtlane jaotus vahemikus $(0; 2)$. Seega

$$p(\nu) = \begin{cases} \frac{1}{2}, \text{ kui } \nu \in (0;2), \\ 0, \text{ mujal .} \end{cases}$$

Tõepärafunktsioon

$$L(\nu) = \prod_{i=1}^{n} f(x_i \mid \nu) = \prod_{i=1}^{n} \nu \exp(-\nu x_i) =$$

$$= \nu^n \exp\left(-\nu \sum_{i=1}^{n} x_i \right) = \nu^n \exp(-\nu n \overline{x}).$$

Rakendades Bayesi teoreemi, saame, et parameetri ν järeljaotus

$$\pi(\nu) \propto \frac{1}{2}\nu^n \exp(-\nu n \overline{x}) \propto \nu^n \exp(-\nu n \overline{x}).$$

4.3.2. Ülevaade Markovi ahelatest

Anname ülevaate Markovi ahelatest. Selleks defineerime esmalt mõiste juhuslik protsess.

Definitsioon 4.12. Juhuslikuks protsessiks nimetatakse ajast t sõltuvat juhuslikku suurust X_t.

Seega kujutab juhuslik protsess endast ajas muutuvat juhuslikku suurust.

Olgu meil juhuslik katse, millel on ülimalt loenduv hulk võimalikke tulemusi $E_1, E_2, ..., E_n, ...$ Katse kordamisel tekib meil juhuslik katsetulemuste loetelu, näiteks E_2, E_3, E_7, E_1, E_1 jne. Iga sellist katsetust võime vaadelda kui ajaühikut. Tegemist on juhusliku protsessiga X_t, mille väärtuste hulk on $E_1, E_2, ..., E_i, ...$ Antud juhul tähistab ajahetke indeks t.

Kui X_t sõltub eelnevatest juhuslikest katsetest üksnes X_{t-1} kaudu, siis on tegemist juhusliku protsessiga, mida nimetatakse Markovi ahelaks.

Definitsioon 4.13. Juhuslike suuruste jada $\{X_t\}$, kus $t = 0, 1, 2, ...$ nimetatakse Markovi ahelaks, kui suvaliste $E_j, E_{k_1}, E_{k_2}, ..., E_{k_{t-1}}$ korral tinglik tõenäosus

$$P(X_t = E_j | X_1 = E_{k_1}, ..., X_{t-2} = E_{k_{t-2}}, X_{t-1} = E_{k_{t-1}}) =$$

$$= P(X_t = E_j | X_{t-1} = E_{k_{t-1}}).$$

Definitsiooni 4.13 iseloomustab järgmine lause: *suvalise oleviku korral tulevik ei sõltu minevikust.* Teisisõnu, Markovi ahel kujutab endast juhuslikku protsessi, kus unustatakse ajalugu. Tegemist on n-ö mäluta protsessiga.

Markovi ahelat iseloomustab üleminekumaatriks

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nn} \end{pmatrix},$$

kus rida tähistab seisundit, milles viibitakse, veerg aga seisundit, kuhu siirdutakse. Iga selle maatriksi element $p_{ij} \leq 1$ ning iga rea puhul

$$\sum_{j=1}^{n} p_{ij} = 1.$$

Maatriksi $\mathbf{P} : n \times n$ elemendiks on tinglik tõenäosus

$$p_{ij} = P(X_t = E_j \mid X_{t-1} = E_i)$$

ehk tõenäosus minna 1 (aja)sammuga seisundist E_i seisundisse E_j. Rea indeks näitab seisundit, kus viibime, ja veeru indeks seisundit, kuhu suundume ühe sammuga. Kui maatriks $\mathbf{P}$ ei sõltu ajast, siis nimetatakse Markovi ahelat homogeenseks. Kuidas aga jõuda homogeense Markovi ahela puhul m sammuga seisundist E_i seisundisse E_j? Vastuse sellele küsimusele annab järgmine seos:

$$p_{ij}(m) = \sum_{l=1}^{n} p_{il}(k) p_{lj}(m-k). \tag{4.16}$$

Seost (4.16) nimetatakse Chapman-Kolmogorovi võrrandiks. Maatrikskujul avaldub Chapman-Kolmogorovi võrrand kui

$$\mathbf{P}(m) = \mathbf{P}(k)\mathbf{P}(m-k),$$

kui $\mathbf{P}(1) = \mathbf{P}$. Seega vastab m sammule üleminekumaatriks

$$\mathbf{P}(m) = \mathbf{P}^m.$$

Olgu seisundite $E_1, E_2, ..., E_n$ jaotus ajahetkel t

$$\pi(t) = \begin{pmatrix} \pi_1(t) & \pi_2(t) & \cdots & \pi_n(t) \end{pmatrix},$$

kus $\sum_{i=1}^{n} \pi_i(t) = 1$. Jaotust $\pi(0)$ nimetatakse Markovi ahela algjaotuseks. Vastavalt Chapman-Kolmogorovi võrrandile saame, et

$$\pi(t+m) = \pi(t)\mathbf{P}^m$$

ehk

$$\pi(1) = \pi(0)\mathbf{P}.$$

Definitsioon 4.14. Seisundite jaotust π^* nimetatakse stationaarseks jaotuseks, kui

$$\pi^* = \pi^*\mathbf{P}.$$

Juhuslikku protsessi (näiteks Markovi ahelat) nimetatakse statsionaarseks, kui tema jaotus ei sõltu ajast. Vastavalt definitsioonile 4.14 tuleb meil statsionaarse jaotuse $\pi = \begin{pmatrix}\pi_1 & \pi_2 & \cdots & \pi_n\end{pmatrix}$ leidmiseks lahendada võrrandisüsteem

$$\pi_i = \sum_{j=1}^{n} \pi_j p_{ij}, \quad i = 1, 2, ..., n,$$

tingimusel, et $\pi_1 + \pi_2 + ... + \pi_n = 1$.

Demonstreerime Markovi ahelate funktsioneerimist mõningate näidete abil.

Näide 4.20. Olgu meil kaks seisundit, mis iseloomustavad siinset ilma: E_1 – sajab ja E_2 – ei saja. Vaheldugu sajupäevad ja sajuta päevad homogeense Markovi ahela seaduspära kohaselt. Olgu üleminekumaatriks

$$\mathbf{P} = \begin{pmatrix}0.7 & 0.3\\ 0.2 & 0.8\end{pmatrix},$$

kus esimene veerg vastab sajupäevadele ja teine sajuta päevadele. Olgu t aeg päevades ning sajupäevade ja sajuta päevade algjaotus

$$\pi(0) = \begin{pmatrix}0.5 & 0.5\end{pmatrix}.$$

Leiame selle jaotuse järgmisel päeval:

$$\pi(1) = \pi(0)P = \begin{pmatrix}0.5 & 0.5\end{pmatrix}\begin{pmatrix}0.7 & 0.3\\ 0.2 & 0.8\end{pmatrix} =$$

$$= \begin{pmatrix} 0.5 \cdot 0.7 + 0.5 \cdot 0.2 & 0.5 \cdot 0.3 + 0.5 \cdot 0.8 \end{pmatrix} = \begin{pmatrix} 0.45 & 0.55 \end{pmatrix}.$$

Milline aga on sajupäevade ja sajuta päevade statsionaarne jaotus? Definitsiooni 4.14 põhjal saame võrrandite süsteemi

$$\begin{cases} 0.7\pi_1 + 0.2\pi_2 = \pi_1 \\ 0.3\pi_1 + 0.8\pi_2 = \pi_2 \end{cases}$$

tingimusel, et $\pi + \pi_2 = 1$. Lahendiks saame, et $\pi_1 = \frac{2}{5}$ ja $\pi_2 = \frac{3}{5}$. Selle põhjal võime öelda, et meil on sajupäevi 40% ja sajuta päevi 60%.

Näide 4.21. Olgu meil elektriskeem, mille eluiga allub eksponentjaotusele, mille parameetrid aga muutuvad ajas. Olgu parameetril 3 erinevat väärtust, mis kajastugu järgnevates seisundites: E_1 – madal töökindlus, E_2 – keskmine töökindlus ja E_3 – kõrge töökindlus. Allugu nende seisundite vaheldumine Markovi ahela eeldustele. Olgu aja sammuks päev ning üleminekumaatriks

$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.2 & 0,6 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}.$$

Leiame töökindluste statsionaarse jaotuse. Selleks lahendame süsteemi

$$\begin{cases} 0.6\pi_1 + 0.2\pi_2 + 0.3\pi_3 = \pi_1 \\ 0.4\pi_1 + 0.6\pi_2 + 0.3\pi_3 = \pi_2 \\ 0.2\pi_2 + 0.7\pi_3 = \pi_3 \end{cases}$$

tingimusel, et $\pi_1 + \pi_2 + \pi_3 = 1$. Saame, et

$$\pi_1 = \frac{2}{7}, \pi_2 = \frac{3}{7} \text{ ja } \pi_3 = \frac{2}{7}.$$

Näide 4.22. Markovi ahelate kolmas näide puudutab hasartmängu. Vaatleme mängijat, kes alustab k rahaühikuga ja kes igas mängus võidab 1 ühiku tõenäosusega p ning kaotab selle ühiku tõenäosusega q. See mängija lõpetab mängu, kui ta jõuab soovitud $n \geq k$ rahaühikuni või ta laostub 0 ühikuga. Tähistagu k rahaühiku olemasolu seisundid $E_0, E_1, ..., E_k, ..., E_n$. Olgu X_m mängija rahaline seis peale m-ndat mängu. Kuna see seis sõltub üksnes seisust X_{m-1} ning m-nda mängu tulemusest, siis on tegemist

Markovi ahelaga. Tema üleminekumaatriks

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ q & 0 & p & 0 & \cdots & 0 \\ 0 & q & 0 & p & \cdots & 0 \\ 0 & 0 & q & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Lahendades seisundite statsionaarse võrrandi (4.14), saame, et $\pi_0 + \pi_n = 1$. Ehk teisisõnu, statsionaarsed saavad olla 2 seisundit:

1) me oleme võitnud soovitud n ühikut;

2) me oleme laostunud 0 ühikuga.

Mängija laostumise probleemist võib huviline lähemalt lugeda õpikust [34] (lk 151–153) või artiklist [33].

Uurime Markovi ahelaid, mille seisundite hulk on mitteloenduv. See tähendab, et seisundite E_i ja E_j asemel on reaalarvud x ja y. Need reaalarvud on juhuslike suuruste X ja Y realisatsioonid. Antud juhul saame üleminekutõenäosuse P_{xy} defineerida järgmiselt:

$$P_{xy} = P(X_{t+1} \leq y \mid X_t = x).$$

Kui P_{xy} on pidev argumendi y suhtes, siis diferentseerides saame tingliku tiheduse

$$p_{xy} = \frac{\partial P_{xy}}{\partial y} = f(y \mid x).$$

Seda tinglikku tihedust nimetatakse Markovi ahela üleminekutuumaks. Ülemineku tõenäosus m sammu korral defineeritakse kui

$$P_{xy}(m) = P(X_{t+m} \leq y \mid X_t = x) = P(X_m \leq y \mid X_0 = x).$$

Seisundite pideval juhul avaldub seos (4.16) kujul

$$P_{xy}(k+m) = \int_{-\infty}^{\infty} p_{xs}(k) p_{sy}(m) ds.$$

Seisundite jaotuse π_t muutuse 1 ajasammuga saame leida järgmiselt:

$$\pi_{t+1}(y) = \int_{-\infty}^{\infty} p_{xy} \pi_t(x) dx.$$

See on analoogne leidmaks juhusliku vektori komponendi üksiktihedust selle vektori ühistiheduse kaudu. Antud juhul kujutavad seisundite jaotused $\pi_t(x)$ ning $p_{t+1}(y)$ endast juhuslike suuruste X ja Y üksiktihedusi. Üleminekutuum p_{xy} aga tinglikku tihedust. Kui jaotus π on statsionaarne, siis

$$\pi(y) = \int_{-\infty}^{\infty} p_{xy}\pi(x)dx.$$

Näide 4.23. Uurime protsessi, mida nimetatakse juhuslikuks ekslemiseks pideval juhul. Sel juhul

$$X_t = X_{t-1} + W_t,$$

kus juhuslikul suurusel W_t on pidev tihedusfunktsioon $f(w)$. Juhuslik suurus X_t avaldub antud juhul sõltumatute juhuslike suuruste (sammude) summana

$$X_t = X_0 + W_1 + W_2 + ... + W_t.$$

Seega on X_t näol tegemist Markovi ahelaga, kus igal ajahetkel liigutakse juhusliku pikkusega sammuga kas vasakule või paremale. Seda vastavalt juhusliku suuruse W_t väärtusele. Ülemineku tõenäosused

$$P_{xy} = P(X_{t+1} \le y \mid X_t = x) = P(X_t + W_{t+1} \le y \mid X_t = x) =$$

$$= P(W_{t+1} \le y - X_t \mid X_t = x) = P(W_{t+1} \le y - x \mid X_t = x) =$$

$$= \int_{-\infty}^{y-x} f(w)dw.$$

Üleminekutuum

$$p_{xy} = \frac{\partial}{\partial y}\int_{-\infty}^{y-x} f(w)dw = f(y-x).$$

Olgu algjaotuseks $\pi(0)$ normaaljaotus $\mathcal{N}(a, \sigma_0)$ ning olgu $W_t \sim \mathcal{N}(0, \sigma)$. Sellisel juhul on antud Markovi ahela (juhusliku ekslemise) jaotuseks $\pi(t)$ normaaljaotus $\mathcal{N}(a, \sqrt{\sigma_0^2 + t\sigma^2})$.

Rahuldagu juhuslik suurus W_t, $t \ge 0$, järgmisi tingimusi.

1° $W_0 = 0$.

2° Iga $t > 0$ korral $W_t \sim \mathcal{N}(0, \sigma\sqrt{t})$, kus $\sigma > 0$ on konstant.

3° Juhusliku suuruse W_t juurdekasvud on sõltumatud ja statsionaarsed (s.t ei sõltu ajahetkest t).

4° Suuruse W_t trajektoorid on t järgi pidevad funktsioonid.

Sel juhul oleme saanud juhusliku protsessi W_t, $t \geq 0$, mida nimetatakse Browni liikumiseks (ehk Wieneri protsessiks).

Viimastel aastakümnetel on Markovi ahelad leidnud üha laialdasemat rakendust statistilises modelleerimises. Need rakendused põhinevad Bayesi teoreemil, mis on seotud parameetrite eel- ja järeljaotustega. Järgnevalt uurime Bayesi statistika ja Markovi ahelate vahelisi seoseid.

4.3.3. Näiteid MCMC-mudelitest

Käsitleme järgnevalt Markovi ahelaid seoses mitmemõõtmeliste jaotuste simuleerimisega. Seda nimetatakse MCMC ehk eesti keeles Markovi ahelad statistikas. Käsitleme simulatsiooni erinevate MCMC-mudelite põhjal. Nendes mudelites on leidnud rakendust juhusliku vektori ning Markovi ahelate teooria tulemused. Markovi ahelate mudelid võib tinglikult jagada kahte gruppi: Gibbsi valik ja Metropolis-Hastingsi algoritm.

Gibbsi valik

Tutvume laialdast rakendust leidnud MCMC simulatsiooni liigiga, mida nimetatakse Gibbsi valikuks (ingl *Gibbs sampling*), kuna see põhineb Gibbsi jaotusel. Selle jaotuse tihedusfunktsioon baseerub keemilise kineetika põhivõrrandil, mis kirjeldab reaktsiooni kiiruse v sõltuvust temperatuurist T

$$v = a \exp\left(- \frac{\Delta E}{RT} \right). \tag{4.17}$$

Seoses (4.17) tähistab R universaalset gaasikonstanti ning ΔE reaktsiooni aktiveerimisenergiat. Selle seose abil saame Gibbsi jaotuse tihedusfunktsiooni avaldada kui

$$f(x_1, x_2, ..., x_k) \propto \exp\left(- \frac{1}{bT} E(x_1, x_2, ..., x_k) \right), \tag{4.18}$$

kus b tähistab positiivset konstanti ning $x_1, x_2, ..., x_k$ on meid huvitava süsteemi karakteristikud. Nendeks karakteristikuteks võivad olla näiteks:

1) füüsikas i-nda osakese asukoht või kiirus;

2) keemias k erinevat keemilist ühendit;

3) kujundi tuvastamise ülesandes oleks i-nda pikseli värviks x_i.

Järgmine definitsioon annab meile Gibbsi valiku olemuse.

Definitsioon 4.15. Gibbsi valikuks nimetatatakse MCMC simulatsiooniskeemi, kus üleminekutuum on moodustatud täistinglike jaotuste abil.

Olgu meid huvitav jaotus $\pi(\Theta)$, kus $\Theta = (\theta_1, \theta_2, ..., \theta_k)^\intercal$. Eesmärk on simuleerida k-mõõtmelist süsteemi Θ jaotuse $\pi(\Theta)$ abil. Kasutame selleks MCMC-meetodit, mida kirjeldab algoritm nimega Gibbsi generaator (ingl *Gibbs generator*). Algoritm ise on järgmine.

1) Algolek, $t = 0$, anname algväärtused $\Theta^0 = (\theta_1^0, \theta_2^0, ..., \theta_k^0)$.

2) Genereerime uue väärtuse $\Theta^t = (\theta_1^t, \theta_2^t, ..., \theta_k^t)$ järgmiste sammudega:

$$\theta_1^t \sim \pi(\theta_1 \mid \theta_2^{t-1}, ..., \theta_k^{t-1});$$

$$\theta_2^t \sim \pi(\theta_2 \mid \theta_1^t, \theta_3^{t-1}, ..., \theta_k^{t-1});$$

$$\vdots$$

$$\theta_k^t \sim \pi(\theta_k \mid \theta_1^t, ..., \theta_{k-1}^t).$$

3) Järgmine samm, $t = t + 1$, kus kordame punkte 2 ja 3.

Antud juhul on tegemist Markovi ahelaga, sest Θ_t saadakse üksnes Θ_{t-1} põhjal. Kui toimub koondamine, siis on Θ_t Markovi ahela statsionaarsest jaotusest π genereeritud väärtus. Põhiline probleem eelkirjeldatud algoritmi juures on täistinglike jaotuste leidmine. Bayesi statistikas on jaotus π meid huvitavate parameetrite järeljaotus.

Vaatame lähemalt näidet, mis pärineb teadusartiklist [4].

Näide 4.24. Olgu meil mõõtmistulemused $\mathbf{y} = (y_1, y_2, ..., y_n)^\intercal$, mis on Poissoni jaotuse mingi realisatsioon. Antud juhul aga kahtlustatakse selle

jaotuse parameetri muutumist alates punktist $m = 1, 2, ..., n$. Selline olukord tekib näiteks juhul, kui me uurime sõltumatute juhuslike sündmuste voogu, mille esimeses osas on üks intensiivsus, teises osas aga teine. Kui m on teada, siis saame kirja panna järgmise mudeli:

$$Y_i \sim Po(\lambda), \quad i = 1, ..., m;$$

$$Y_i \sim Po(\phi), \quad i = m+1, ..., n.$$

Fikseerime antud Poissoni jaotuse parameetrite eeljaotused, nagu on nõutud Bayesi statistikas. Olgu selleks järgmised jaotused:

$$\lambda \sim G(\alpha, \beta);$$

$$\phi \sim G(\gamma, \delta);$$

$$P(m = i) = \frac{1}{n}, \quad i = 1, 2, ..., n.$$

Eeldame, et parameetrid λ, ϕ ja m on sõltumatud juhuslikud suurused ning parameetrid α, β, γ ja δ on teadaolevad konstandid. Siis saame Bayesi teoreemi põhjal järeljaotuseks

$$\pi(\lambda, \phi, m) \propto f(y_1, y_2, ..., y_n \mid \lambda, \phi, m) p(\lambda, \phi, m) =$$

$$= \left(\prod_{i=1}^{m} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} \right) \left(\prod_{i=m+1}^{n} \frac{\phi^{y_i}}{y_i!} e^{-\phi} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \frac{\delta^\gamma}{\Gamma(\gamma)} \phi^{\gamma-1} e^{-\delta\phi} \frac{1}{n} \propto$$

$$\propto \left(\prod_{i=1}^{m} e^{-\lambda} \lambda^{y_i} \right) \left(\prod_{i=m+1}^{n} e^{-\phi} \phi^{y_i} \right) \lambda^{\alpha-1} e^{-\beta\lambda} \phi^{\gamma-1} e^{-\delta\phi} =$$

$$= \lambda^{\alpha + \sum_{i=1}^{m} y_i - 1} \exp(-(\beta + m)\lambda) \phi^{\gamma + \sum_{i=m+1}^{n} y_i - 1} \exp(-(\delta + n - m)\phi).$$

Olgu $\pi_1(\lambda)$ parameetri λ täistinglik jaotus ning $\pi_2(\phi)$ parameetri ϕ täistinglik jaotus. Siis korjates kokku kõik suurusi λ ja ϕ sisaldavad liikmed ning jättes ülejäänud liikmed konstandiks saame, et

$$\pi_1(\lambda) = \frac{(\beta + m)^{\alpha + \sum_{i=1}^{m} y_i}}{\Gamma\left(\alpha + \sum_{i=1}^{m} y_i\right)} \lambda^{\alpha + \sum_{i=1}^{m} y_i - 1} \exp(-(\beta + m)\lambda);$$

$$\pi_2(\phi) = \frac{(\delta + n - m)^{\gamma + \sum_{i=m+1}^{n} y_i}}{\Gamma\left(\gamma + \sum_{i=m+1}^{n} y_i\right)} \phi^{\gamma + \sum_{i=m+1}^{n} y_i - 1} \exp(-(\delta + n - m)\phi).$$

Seega saime parameetri λ järeljaotuseks gammajaotuse $G(\alpha+\sum_{i=1}^{m} y_i, \beta+m)$ ning parameetri ϕ järeljaotuseks gammajaotuse $G(\gamma+\sum_{i=m+1}^{n} y_i, \delta+n-m)$. Parameetri m täistinglik jaotus

$$\pi_3(m) = \pi_3(m \mid \lambda, \phi) = \frac{\pi(\lambda, \phi, m)}{\pi(\lambda, \phi)},$$

kus $\pi(\lambda, \phi) = \sum_{m=1}^{n} \pi(\lambda, \phi, m)$. Korjates järeljaotuses $\pi(\lambda, \phi, m)$ kokku kõik suurusest m sõltuvad liikmed, saame, et

$$\pi_3(m) \propto \frac{\lambda^{\sum_{i=1}^{m} y_i} \phi^{\sum_{i=m+1}^{n} y_i} \exp(m(\phi-\lambda))}{\sum_{m=1}^{n} \lambda^{\sum_{i=1}^{m}} \phi^{\sum_{i=m+1}^{n}} \exp(m(\phi-\lambda))} =$$

$$= \frac{\lambda^{\sum_{i=1}^{m} y_i} \phi^{\sum_{i=m+1}^{n} y_i}}{\lambda^{y_1} \phi^{\sum_{i=2}^{n} y_i} \exp(\phi-\lambda) + \ldots + \lambda^{\sum_{i=1}^{n} y_i} \exp(n(\phi-\lambda))} \exp(m(\phi-\lambda)).$$

Jagades lugeja ning nimetaja suurusega $\phi^{\sum_{i=1}^{m} y_i} \phi^{\sum_{i=m+1}^{m} y_i}$, saame, et

$$\pi_3(m) \propto \left(\frac{\lambda}{\phi}\right)^{\sum_{i=1}^{m} y_i} \exp(m(\phi-\lambda)).$$

Leitud jaotustest saame teostada Gibbsi valiku sammud 1–3. Olgu

$$\Theta^0 = (\lambda^0, \phi^0, m^0).$$

Võtame $m^0 = 1$. Gammajaotuse keskväärtust arvestades olgu

$$\lambda^0 = \frac{\alpha}{\beta} \text{ ning } \phi^0 = \frac{\gamma}{\delta}.$$

Järgnevate sammudena saame realisatsioonid järeljaotustest

$$\Theta^t = (\lambda^t, \phi^t, m^t) \sim \pi(\lambda, \phi, m).$$

Simulatsiooni korratakse niikaua, kuni tekib statsionaarne jaotus $\pi^*(\lambda, \phi, m)$.

Näide 4.25. Olgu meil juhuslik vektor $\theta = (\theta_1, \theta_2)^\intercal$, mille komponendid θ_i allugu Bernoulli jaotusele, $i = 1, 2$. Olgu vektorit θ iseloomustav seisundite ruum $\{0, 1\}^2$. Selle vektori jaotust kirjeldagu järgmine sagedustabel:

$\theta_1 \setminus \theta_2$	0	1
0	π_{00}	π_{01}
1	π_{10}	π_{11}

On selge, et $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1$. Vastavalt Gibbsi valiku reeglitele saame järgmised üleminekutõenäosused.

1) Komponendi θ_1 puhul saame tinglikeks tõenäosusteks

$$\pi_1(0 \mid j) = \frac{\pi_{0j}}{\pi_{0j} + \pi_{1j}} \text{ ning } \pi_1(1 \mid j) = \frac{\pi_{1j}}{\pi_{0j} + \pi_{1j}},$$

kus $i = 0, 1$.

2) Komponendi θ_2 puhul saame tinglikeks tõenäosusteks

$$\pi_2(0 \mid i) = \frac{\pi_{i0}}{\pi_{i0} + \pi_{i1}} \text{ ning } \pi_2(1 \mid i) = \frac{\pi_{i1}}{\pi_{i0} + \pi_{i1}},$$

kus $i = 0, 1$.

Kogu ahela puhul saame üleminekutõenäosuseks

$$P((i,j)(k,l)) = P(\theta^t = (k,l) \mid \theta^{t-1} = (i,j)) =$$
$$= P(\theta_2^t = l \mid \theta_1^t = k) P(\theta_1^t = k \mid \theta_2^{t-1} = j) = \frac{\pi_{kl}\pi_{kj}}{(\pi_{k0} + \pi_{k1})(\pi_{0j} + \pi_{1j})},$$

kus $(i,j)(k,l) \in \{0,1\}^2$. Selliselt saame leida 4×4-maatriksi $\mathbf{P}$, mis on Markovi ahela üleminekumaatriks. Selle Markovi ahela seisunditeks on vektori θ võimalikud väärtused

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ ja } \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Metropolis-Hastingsi algoritm

Uurime Markovi ahelate mudeleid kontekstis, mida nimetatakse Metropolis-Hastingsi algoritmiks. Seda algoritmi rakendati esmakordselt 1953. aastal arvutamaks keemiliste ühendite konfiguratsiooni muutusi. Need arvutused on publitseeritud teadusartiklis [30].

Olgu meil k mingi aine molekuli konfiguratsiooniga $\theta = (\theta_1, \theta_2, ..., \theta_k)^\intercal$. Alternatiivseks positsiooniks olgu $\phi = (\phi_1, \phi_2, ..., \phi_k)^\intercal$. Näiteks glükoosi molekuli puhul oleksid need konfiguratsioonid nimedega „tugitool“ ja „vann“. Kuidas arvutada üleminekut konfiguratsioonist θ konfiguratsiooni ϕ? Artiklis [30] soovitatakse järgmist meetodit modelleerimaks keeruliste keemiliste süsteemide muutusi.

1) Olgu alghetkel keemiliste ühendite konfiguratsioon $\theta^0 = (\theta_1^0, \theta_2^0, ..., \theta_k^0)^\intercal$.

2) Allugu üleminek konfiguratsioonist $\theta^{j-1} = (\theta_1^{j-1}, \theta_2^{j-1}, ..., \theta_k^{j-1})^\intercal$ uude konfiguratsiooni $\phi = (\phi_1, \phi_2, ..., \phi_k)^\intercal$ ühtlasele jaotusele.

3) Leida konfiguratsiooni muutusest tekkinud energia muut ΔE. Aktsepteerida eelmises punktis kirjeldatud konfiguratsiooni muut tõenäosusega

$$\alpha_{\theta\phi} = \min\{1, \exp(-c\Delta E)\},$$

kus $c = \dfrac{1}{bT}$, milles b tähistab Boltzmanni konstanti ning T temperatuuri absoluutsel skaalas. Kui muutus toimus, siis $\theta^j = \phi$, vastasel juhul $\theta^j = \theta^{j-1}$.

4) Suurusele j omistatakse suurus $j + 1$ ning minnakse tagasi punkti 2, kuni on saavutatud koondumine.

Pärast algortmi koondumist on konfiguratsioonide jaotust kirjeldav tihedusfunktsioon kujuga (4.18). Kuna eelpool kirjeldatud algoritmi puhul sõltub üleminek järgmisse seisundisse (positsiooni) üksnes eelmisest seisundist, siis on tegemist Markovi ahelaga. Antud ajahetke ja eelmise ajahetke seisundi jaotuste suhe avaldub kui

$$\frac{\pi(\phi)}{\pi(\theta^{j-1})} = \frac{\exp(-cE(\phi))}{\exp(-cE(\theta^{j-1}))} =$$

$$= \exp\{-c(E(\phi) - E(\theta^{j-1}))\} = \exp(-c\Delta E).$$

Seega energiabarjääri ΔE kasvades konfiguratsiooni muutuse aktsepteerimise tõenäosus kahaneb. Teisisõnu, mida suurem on uue konfiguratsiooni ϕ potentsiaalne energia võrreldes endise konfiguratsiooni θ omaga, seda suurema tõenäosusga jääb konfiguratsiooni muutus olemata.

Uurime lähemalt Metropolis-Hastingsi algoritmi olemust. Me tahame genereerida jaotuse π Markovi ahela abil. Otsese ülemineku põhjal on se-

da teha väga raske või isegi võimatu. Meil tuleb konstrueerida üleminekutuum $p_{\theta\phi}$, mis kujutab endast tinglikku tihedust. Peame leidma sellise jaotuse π, et kehtiks seos

$$\pi(\theta)p_{\theta\phi} = \pi(\phi)p_{\phi\theta}. \tag{4.19}$$

Võrrandit (4.19) nimetatakse detailse tasakaalu võrrandiks. Selle võrrandi kehtides on jaotus π tasakaalu jaotus ehk piirjaotus.

Üleminekutuum sisaldab 2 elementi

$$p_{\theta\phi} = q_{\theta\phi}\alpha_{\theta,\phi}. \tag{4.20}$$

Element $q_{\theta\phi}$ tähendab suvalist üleminekutuuma, elementi $\alpha_{\theta,\phi}$ aga nimetatakse vastuvõtmise tõenäosuseks ehk ettepaneku tuumaks. Kui $\theta = \phi$, siis

$$p_{\theta\theta} = 1 - \int q_{\theta\phi}\alpha_{\theta\phi}d\phi. \tag{4.21}$$

Seos (4.21) annab meile tõenäosuse jääda samasse seisundisse. Artiklis [17] tegi Hastings ettepaneku leidmaks vastuvõtmise tõenäosust järgmise eeskirjaga:

$$\alpha_{\theta\phi} = \min\left\{1, \frac{\pi(\phi)q_{\phi\theta}}{\pi(\theta)q_{\theta\phi}}\right\}. \tag{4.22}$$

Kui $\pi(\phi)q_{\phi\theta} \geq \pi(\theta)q_{\theta\phi}$, siis võetakse muutus seisundist θ seisundisse ϕ vastu tõenäosusega 1.

Definitsioon 4.16. Algoritmi üleminekutuumadega (4.20)-(4.21) ning vastuvõtmise tõenäosusega (4.22) nimetatakse Metropolis-Hastingsi algoritmiks.

Metropolis-Hastingsi algoritmi on järgmine.

1) Olgu ajahetkel $j = 0$ algväärtuseks θ^0.

2) Genereerime uue väärtuse ϕ tõenäosustiheduse $q(\theta_{j-1})$ põhjal.

3) Leiame avaldise (4.22) põhjal vastuvõtmise tõenäosuse $\alpha(\theta^{j-1}, \phi)$.

4) Genereerime juhusliku suuruse U, mis allub ühtlasele jaotusele lõigul $[0; 1]$. Kui selle juhusliku suuruse väärtus $u < \alpha(\theta^{j-1}, \phi)$, siis võtame vastu muutuse ehk $\theta^j = \phi$. Vastasel korral jääb muutus vastu võtmata ehk $\theta^j = \theta^{j-1}$.

5) Omistame $j := j+1$ ning siirdume algoritmi teise punkti juurde, kuni on saavutatud koondumine.

Metropolis-Hastingsi algoritmi juures tekib küsimus, kuidas hinnata vastuvõtmise tõenäosuse $\alpha_{\theta,\phi}$ headust. See tõenäosus kujutab endast vastuvõetud otsuste hulga ja iteratsioonide hulga suhet. See suhe ei tohi olla liiga väike ega ka liiga suur. Rusikareegli kohaselt loetakse vastuvõtu tõenäosust heaks, kui see jääb 0.2 ja 0.5 vahele. Vaatame kahte äärmuslikku juhtu.

1) Kui ϕ ja θ on üksteisele lähedal, siis $\alpha_{\theta,\phi} \approx 1$. Sel juhul on $\Delta\theta$ absoluutväärtus näites 4.26 väike. Enamus ettepanekuid parameetri muutmiseks võetakse vastu, aga kiirus on aeglane.

2) Kui ϕ ja θ on teineteisest kaugel, siis on vastuvõtmise tõenäosus $\alpha_{\theta,\phi}$ väike. Nüüd on näites 4.26 $\Delta\theta$ suur. Kandja kiirus on antud juhul küll suur, kuid enamus parameetri muutmise ettepanekutest lükatakse tagasi.

Demonstreerime Metropolis-Hastingsi algoritmi ühenduses Bayesi teoreemiga. Selle teoreemi kohaselt on jaotustihedus $\pi(\theta)$ proportsionaalne tõepära funktsiooni ning eeljaotuse tihedusfunktsiooni korrutisega ehk

$$\pi(\theta) \propto l(\theta)p(\theta).$$

Vaatame juhtu, kus vastuvõtmise tõenäosus sõltub tõepära suhtest.

Näide 4.26. Mingit sündmust A katsetatakse sõltumatult n korda. Olgu juhuslik suurus X õnnestunud katsete hulk ning tõenäosus $P(A) = \theta$. Siis

$$P(X = x) = \theta^x(1-\theta)^{n-x}.$$

Parameetri θ eeljaotuseks olgu beetajaotus $Beta(\alpha, \beta)$. Rakendame sündmuse A tõenäosusele θ Metropolis-Hastingsi algoritmi. Allugu muut $\Delta\theta$ normaaljaotusele keskväärtusega 0 ja standardhälbega σ. Olgu esialgne parameeter θ_c ning kavandatav parameeter $\theta_p = \theta_c + \Delta\theta$. Siis saame kirja panna järgmise algoritmi.

1) Genereerime juhuslike arvude generaatoriga $\Delta\theta \sim \mathcal{N}(0, \sigma)$. Võtame kavandatavaks parameetriks $\theta_p = \theta_c + \Delta\theta$.

2) Moodustame kavandatava parameetri vastuvõtmise tõenäosuse

$$\alpha(\theta_c, \theta_p) = \min\left\{1, \frac{P(\theta_p)}{P(\theta_c)}\right\} = \min\left\{1, \frac{P(X = x \mid \theta_p)p(\theta_p)}{P(X = x \mid \theta_c)p(\theta_c)}\right\} =$$

$$= \min\left\{1, \frac{\theta_p^x(1-\theta_p)^{n-x}Beta(\theta_p \mid \alpha, \beta)}{\theta_c^x(1-\theta_c)^{n-x}Beta(\theta_c \mid \alpha, \beta)}\right\} =$$

$$= \min\left\{1, \frac{\theta_p^x(1-\theta_p)^{n-x}\theta_p^{\alpha-1}(1-\theta_p)^{\beta-1}}{\theta_c^x(1-\theta_c)^{n-x}\theta_c^{\alpha-1}(1-\theta_c)^{\beta-1}}\right\}.$$

3) Genereerime juhuslike arvude generaatoriga lõigul $[0;1]$ ühtlase jaotusega juhusliku suuruse U. Aktsepteerime parameetri väärtust θ_p, kui $u < \alpha(\theta_c, \theta_p)$. Vastasel korral lükkame kavandatud väärtuse tagasi ning jätkame väärtusega θ_c.

Näites 4.26 toodud algoritmile saab leida mitmeid rakendusi:

1) populatsioonigeneetikas saab sellega uurida mingi geeni alleeli sageduse dünaamikat;

2) sündmuste voo intensiivsuse ajalise muutuse modelleerimisel;

3) samuti saab selle näite algoritmiga modelleerida metsastuse osakaalu muutust aastate lõikes.

Järgnevalt rakendame Metropolis-Hastingsi algoritmi uurimaks raha teenimise ja kulutamise dünaamikat.

Näide 4.27. Olgu meil alghetkel mingi rahasumma v_0. Vaatame seda alghetke kui nullseisu ehk $v_0 = 0$. Allugu raha muut Δv ühtlasele jaotusele lõigus $[-a; a]$. Olgu üleminekutuumaks standardsele normaaljaotusele vastav tõenäosus. Nende eelduste korral saame rahahulga muutuse vastuvõtmise kriteeriumiks

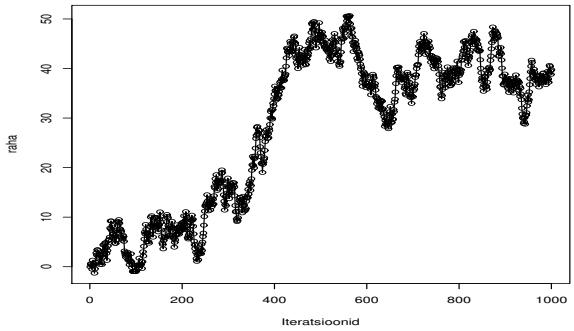
$$\alpha(v_i, v_{i+1}) = \min\left\{1, \frac{\Phi(v_i + \Delta v) + 0.5}{\Phi(v_i) + 0.5}\right\}, i = 0, 1, ..., n.$$

Genereerime taas juhusliku suuruse U, mis allub ühtlasele jaotusele lõigus $[0;1]$. Kui $u < \alpha(v_i, v_{i+1})$, siis $v_{i+1} = v_i + \Delta v$. Vastasel juhul $v_{i+1} = v_i$.

Tarkvaras R saab seda algoritmi realiseerida järgmiselt:

```
a=2
kesk=0
shalve=1
raha=vector("numeric", 1000)
```

```
x=0
raha[1]=x
for (i in 2:1000) {
muut=runif(1, -a, a)
uus=x+muut
alpha=min(1, pnorm(uus,kesk,shalve)/pnorm(x,kesk,shalve))
u=runif(1)
if (u < alpha)
{x=uus} else {x=x}
raha[i]=x}.
```



Joonis 4.3. Raha muutus ajas parameetri $a = 2$ korral

Joonisel 4.3 on toodud raha muutumise üks võimalikest dünaamikatest. Näite 4.27 puhul on tegemist juhuga, mille puhul rahahulk pigem kasvab, kuid eksisteerib risk oma kasum maha mängida. Mida suurem on rahahulk, seda suuremaks muutub ka raha kulutamine.

4.3.4. Tarkvara OpenBUGS

Teeme tutvust ühe Bayesi meetodite ja MCMC tarkvaraga. See põhineb 1989. aastal alguse saanud projektil nimega BUGS (ingl ***B****ayesian inference* ***U****sing* ***G****ibbs* ***S****ampling*). Just sel ajal seoti Gibbsi valik Bayesi statistikaga. Tarkvara OpenBUGS on vabavara, millega saab töötada nii Windowsi kui ka Linuxi keskkonnas.

1) Windowsi keskkonnas töötamiseks saab OpenBUGSi tarkvara alla laadida lingilt
`http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/`.

2) Keskkonnas Linux saab seda teha lingilt
`http://nd.psychstat.org/blog/installl_openbugs_3.0.7_on_linux_server`.

Esmalt anname juhised, kuidas selle tarkvaraga töötada. Seejärel tutvume OpenBUGSi baasil mõningate konkreetsete mudelitega.

Põhikäsud tarkvaraga OpenBUGS töötamisel

Tarkvaraga OpenBUGS käib kaasas suur hulk MCMC-mudeleid. Alljärgnevalt punktide kaupa nendega töötamise reeglid.

1) Esimeses etapis valime käsuaknast *File* käsuga *Open* soovitud mudeli kataloogist *Examples*.

2) Teist etappi võiks nimetada kui mudeli töökorda seadmist. Valime käsuaknast *Model* käsu *Specification Tool*. Seepeale ilmub ekraanile aken nimega *Specification Tool*. Märgistame kirjutatud mudeli hiirega ja valime käsu *Check model*. Vastuseks peame saama, et mudel on süntaksi mõttes korrektne. Seejärel märgistame andmed ning loeme need sisse käsuga *Load data*. Siis kompileerime mudeli käsuga *Compile*. Lõpuks loeme sisse mudeli parameetrite algväärtused käskudega *Load inits* ja *Gen inits*.

3) Kolmandat etappi võiks nimetada kui parameetrite modelleerimist. Valime käsuaknast *Inference* käsu *Samples*. Seejärel ilmub kuvari ekraanile aken nimega *Sample Monitor Tool*. Selles aknas koostame loetelu meid huvitavatest parameetritest. Pärast seda valime käsuaknast *Model* käsu *Update*. Seejärel avaneb meile aken nimega *Update Tool*. Selles aknas valime Markovi ahela sammude arvu.

Pärast ülaltoodud 3 etappi peaks olema meid huvitav mudel valmis analüüsimiseks. Selleks valime taas käsuaknast *Inference* käsu *Samples*. Siis kirjutame aknasse *Node* meid huvitava parameetri. Seepeale ilmuvad selle parameetri analüüsimiseks erinevad võimalused: *density*, *history*, *quantiles* jms. Nende võimalustega saame uurida meid huvitavate parameetrite

järeljaotuseid ning leida nende α-usaldusintervalle.

Näiteid tarkvara OpenBUGS mudelitest

Tutvume lähemalt tarkvara OpenBUGS kolme mudeliga. Uurime nende mudelite sisendeid ja väljundeid ehk eel- ja järeljaotuseid. Selle uurimise käigus teeme tutvust tarkvara süntaksiga.

Pumpade mudel (ingl *Pumps*)

Pumpade mudelit on põhjalikumalt kirjeldatud teadusartiklis [13]. Näide puudutab 10 elektrijaama pumpa. Iga pumba juures on uuritud selle tõrgete sagedust $X_i \sim Po(\theta_i t_i)$, $i = 1, 2, ..., 10$. Suurus θ_i iseloomustab i-pumba töö intensiivsust (tõrgete hulk tunnis) ning t_i selle pumba tööaja kestvust tundides. Varasema info põhjal on intensiivsuse θ_i eeljaotuseks võetud gammajaotus $G(\alpha, \beta)$. Selle jaotuse parameetrite puhul eeldatakse, et

$$\alpha \sim \mathcal{E}(1) \text{ ning } \beta \sim G(0.1, 1).$$

Neid eeldusi korrigeeriti järgmise andmestikuga:

Pump	t_i	x_i
1	94.5	5
2	15.7	1
3	62.9	5
4	126	14
5	5.24	3
6	31.4	19
7	1.05	1
8	1.05	1
9	2.1	4
10	10.5	22

Andmestiku alla laadimiseks tuleb see sisestada kujul

```
list(t = c(94.3, 15.7, 62.9, 126, 5.24, 31.4, 1.05, 1.05, 2.1,
10.5),
x = c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22), N = 10).
```

Alljärgnevalt on kirja pandud mudel OpenBUGSi tarkvara keeles:

```
model
{
for (i in 1 : N) {
theta[i] ∼ dgamma(alpha, beta)
lambda[i] <- theta[i] * t[i]
x[i] ∼ dpois(lambda[i])
}
alpha ∼ dexp(1)
beta ∼ dgamma(0.1, 1.0)
}.
```

Enne mudeli kompileerimist loeti järgmiselt sisse parameetrite α ja β algväärtused:

```
list(alpha = 1, beta = 1).
```

Pärast 1000 iteratsiooni sammu saadi parameetritele α ja β järeljaotused, mille põhjal leiti neile karakteristikud. Mõningad neist karakteristikutest on toodud järgmises tabelis:

Parameeter	Keskmine	Mediaan	Standardhälve
α	0.743	0.687	0.417
β	0.996	0.86	0.813

Nende karakteristikute abil saadi intensiivsustele θ_i, $i = 1, 2, ..., 10$, järeljaotused, mille põhjal leiti järgmised 0.95-usaldusintervallid:

Intensiivsus	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
Alumine	0.0229	$7.66 \cdot 10^{-3}$	0.032	0.0636	0.151	0.372	0.0685	0.0229	0.435	1.27
Ülemine	0.119	0.302	0.169	0.183	1.32	0.909	2.77	0.119	3.65	2.95

Õhusaaste mudel (ingl *Air*)

Selles mudelis rakendatakse Bayesi meetodeid uurimaks laste haigestumise riski teatud hingamisteede haigustesse sõltuvalt lämmastikdioksiidi (NO_2) hulgast. Uuriti 103 last. Mõõdeti lämmastikdioksiidi hulka (μg/l) lapse magamistoas. See hulk jagati 3 kontsentratsiooni klassi: < 20, $20 - 40$ ning $40 +$. Saadi järgmine sagedustabel:

Haigestumine Y	< 20	20–40	40+	Kokku
Jah	21	20	15	56
Ei	27	14	6	47
Kokku	48	34	21	103

Kontsentratsiooni klassi z_j ning tegelikult manustatud NO_2 hulga x_j vahel on leitud seos

$$x_j = \alpha + \beta z_j + \epsilon_j,$$

kus $\alpha = 4.48$, $\beta = 0.76$ ning juhuslik suurus ϵ_j allub normaaljaotusele $\mathcal{N}(0, 9.01)$ $j = 1, 2, 3$. Uuritavale tunnusele Y koostati mudel

$$Y_j \sim B(p_j, n_j) \text{ ning } p_j = \frac{\exp(\theta_1 + \theta_2 x_j)}{1 + \exp(\theta_1 + \theta_2 x_j)},$$

kus p_j iseloomustab haigestumise tõenäosust j-nda kontsentratsiooni taseme korral ning $\theta_1 \sim \mathcal{N}(0, \sqrt{0.001})$ ja $\theta_2 \sim \mathcal{N}(0, \sqrt{0.001})$. Andmed laaditi alla ning loeti sisse käsuga

```
list(J = 3, y = c(21, 20, 15), n = c(48, 34, 21),
Z = c(10, 30, 50), tau = 0.01234, alpha = 4.48, beta = 0.76).
```

Mudeli süntaks on järgmine:

```
model
{
for(j in 1 : J)
{y[j] ~ dbin(p[j], n[j])
logit(p[j]) <- theta[1] + theta[2]*X[j]
X[j] ~ dnorm(mu[j], tau)
mu[j] <- alpha + beta*Z[j]
}
theta[1] ~ dnorm(0.0, 0.001)
theta[2] ~ dnorm(0.0, 0.001)
}.
```

Algväärtused loeti sisse järgmiselt:

```
list(theta = c(0.0, 0.0), X = c(0.0, 0.0, 0.0)).
```

Pärast 1000 iteratsiooni sammu saadi haigestumise tõenäosuste 0.95-usaldusintervallidele järgmised alumised ja ülemised piirid:

NO_2 hulk (μg/l)	Alumine	Ülemine
< 20	0.32	0.574
20–40	0.445	0.724
40+	0.508	0.866

Lineaarse regressiooni mudel

Rakendame Bayesi statistikat üldistele lineaarsetele mudelitele. Mudel tervikuna on järgmine:

$$Y_i \sim \mathcal{N}(\mu_i, \tau) \text{ ja } \mu_i = \alpha + \beta(x_i - x^0), \quad i = 1, 2, ..., n.$$

Hajuvust iseloomustava parameetri τ eeljaotuseks on gammajaotus

$$G(0.001, 0.001).$$

Lineaarse mudeli parameetrite α ja β eeljaotuseks on võetud normaaljaotus $\mathcal{N}(0, 10^{-3})$. Tarkvara OpenBUGS süntaksis on mudel järgmise kujuga:

```
model
{
for( i in 1 : N ) {
Y[i] ~ dnorm(mu[i],tau)
mu[i] <- alpha + beta*(x[i] - xbar)
}
tau ~ dgamma(0.001, 0.001) sigma <- 1 / sqrt(tau)
alpha ~ dnorm(0.0,1.0E-6)
beta ~ dnorm(0.0,1.0E-6)
}.
```

Sisse loetav andmestik on järgmine:

```
list(x = c(1, 2, 3, 4, 5), Y= c(1, 3, 3, 3, 5),
xbar = 3, N = 5).
```

Seega on antud juhul valimi maht $n = 5$. Parameetritele α, β ja τ anti järgmised algväärtused:

```
list(alpha = 0, beta = 0, tau = 1).
```

Pärast 1000 iteratsiooni sammu saadi mudeli parameetritele α ja β ning hajuvuse parameetrile τ järgmised 0.95-usaldusintervallid:

Parameeter	Alumine	Ülemine
α	2.056	4.073
β	0.1231	1.461
τ	0.1847	5.795

Kui uurida eespool toodud 3 mudeli programme üldisemalt, siis võib täheldada, et need koosnevad kahest osast:

1) andmestiku osast (ehk *for* tsükli sees olevast osast), milles on kirjeldatud andmestikus esinevate tunnuste jaotused;

2) eeljaotuste osast, kus on toodud nende jaotuste parameetrite teoreetilised jaotused.

Mudelite väljunditeks on parameetrite jaotuste järeljaotused.

Märkus. Programme MCMC-mudelitele võib kirjutada Notepadi keskkonnas ning failid salvestada .txt formaadis. Seejärel saab neid faile kopeerida OpenBUGSi keskkonda.

Seosed tarkvarade R ja OpenBUGS vahel

Tarkvara OpenBUGS mudelitega saab töötada ka R-i keskkonnas. Selleks tuleb installeerida tarkvara R pakett `R2OpenBUGS`. Enne seda tuleb alla laadida fail `R2OpenBUGS_3.2-3.1.tar.gz`. Installeerimine toimub käsuga

`install.packages("R2OpenBUGS",lib="faili nimi").`

Argumendiga `lib` tuleb määrata faili `R2OpenBUGS_3.2-3.1.tar.gz` asukoht.

Demonstreerime keskkonnas R modelleerimist lineaarse regressiooni mudeli baasil. Uurime seda etappide kaupa.

1) Kõigepealt tuleb sisse lugeda andmed:

```
x=c(1,2,3,4,5)
Y=c(1,3,3,3,5).
```

2) Seejärel tuleb andmed ette valmistada:

```
andmed=cbind(x,Y)
N=nrow(andmed)
xbar=3
data=list("x","Y","xbar","N").
```

3) Järgnevalt tuleb anda mudeli parameetritele algväärtused:

```
inits=function(){list(alpha=rnorm(1,0,1.0e-6),
beta=rnorm(1,0,1.0e-6),tau=rgamma(1,0.001,0.001))}.
```

4) Viimaseks etapiks on iteratsioonide läbi viimine:

```
lin.sim=bugs(data,inits,model.file="faili nimi",
parameters=c(”alpha","beta","tau"),
n.chains=2,n.iter=1000,debug=TRUE).
```

Argumendiga `model.file` tuleb määrata OpenBUGSi süntaksiga koostatud mudeli täpne asukoht. Mudeli programmi fail peab olema salvestatud .txt formaadis. Mudeli väljundi saab tellida käsuga

```
print(lin.sim).
```

Selle käsuga saame antud juhul tabeli, kus on toodud lineaarse regressiooni mudeli parameetrite α, β ja τ keskväärtused, standardhälbed ja erinevad kvantiilid.

Graafilisel kujul saab esitada neid parameetreid käsuga

```
plot(lin.sim).
```

Tarkvara OpenBUGS edasiarendus

Viimasel ajal on koostatud eri valdkondades (molekulaarbioloogias, metsanduses, ökoloogias) üha keerulisemaid MCMC-mudeleid. Need mudelid nõuavad tarkvara OpenBUGS edasiarendusi. Üks selline arendus on tarkvara Stan. Oma nime on Stan saanud Monte Carlo meetodite ühe rajaja Stanislaw Ulami järgi. Ka Stani näol on tegemist vabavaraga. Oma olemuselt on Stan C++ programmeerimiskeel, mis on kirjutatud statistiliste mudelite jaoks. See on sobilik töövahend keeruliste hierarhiliste mudelite juures, kus OpenBUGS jääb sageli hätta.

Juhised installeerimaks tarkvara Stan saab järgmistelt linkidelt:

1) keskkonna Windows korral `https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Windows`;

2) keskkonna Linux puhul `https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Mac-or-Linux`.

Stan liidestub tarkvaraga R järgmise käsuga:

```
install.packages("rstan",
repos = "https://cloud.r-project.org/", dependencies=TRUE).
```

Pärast seda käsku saab Stani programmidega töötada R-i keskkonnas.

Demonstreerime lihtsa näite abil Stani süntaksit.

Näide 4.28. Olgu meil uuritav tunnus Y, mis võtab kas väärtuse 0 või 1. Olgu

$$\theta = P(Y = 1),$$

mis on parameetritega 1 ja 0.5 beetajaotuse realisatsioon. Seega on eeljaotuseks jaotus $Beta(1, 0.5)$. Mudeli süntaks on järgmine:

```
mudel='
data {
int<lower=0> N;
int<lower=0,upper=1> y[N];
}
parameters {
real<lower=0,upper=1> theta;
}
model {
theta~beta(1,0.5);
y~bernoulli(theta);
}
'.
```

Andmed ja mudel loetakse sisse järgmiselt:

```
andmed=list(N=10,y=c(1,0,0,0,1,1,1,1,0,1))
valjund=stan(model_code=mudel,data=andmed,iter=500,chains=1).
```

Käskudega

```
print(valjund)
plot(valjund)
```

saadakse mudeli väljundid tabelina ning graafiliselt.

Põhjalikumat infot tarkvara Stan ja tema süntaksi kohta võib huviline leida artiklist [5].

4.4. Ülesanded

Ülesanne 4.1. Iseloomustagu tunnus X inimese haridust (0 – alghariduse, 1 – keskhariduse ja 2 – kõrghariduse) ning tunnus Y tema suhtumist mingisse erakonda (-1 – negatiivne, 0 – neutraalne ja 1 – positiivne). Küsitleti 100 elanikku ja saadi järgmine sagedustabel:

X \ Y	-1	0	1
0	12	8	8
1	4	16	6
2	8	20	18

Leidke regressioonikordajad $\gamma(X,Y)$ ja $\gamma(Y,X)$ ning χ^2-kordaja $\chi(X,Y)$. Mida nende põhjal järeldada?

Ülesanne 4.2. Olgu meil juhuslik suurus $X = 0, 1, 2$, mille jaotus on esitatud tabelina

x_i	-1	0	1
p_i	0.3	0.3	0.4

Olgu juhuslik suurus $Y = X^2$ ning juhuslik suurus $Z = 2^Y$. Leidke juhuslike vektorite $\mathbf{X}_1 = (X,Y)^\intercal$, $\mathbf{X}_2 = (X,Z)^\intercal$ ja $\mathbf{X}_3 = (Y,Z)^\intercal$ ühisjaotused ja lineaarsed korrelatsioonikordajad $\mathrm{corr}(X,Y)$, $\mathrm{corr}(X,Z)$ ning $\mathrm{corr}(Y,Z)$. Seejärel leidke regressioonikordajad

1) $\gamma(X,Y)$ ja $\gamma(Y,X)$;

2) $\gamma(X,Z)$ ja $\gamma(Z,X)$;

3) $\gamma(Y,Z)$ ja $\gamma(Z,Y)$.

Ülesanne 4.3. Olgu juhusliku vektori $\mathbf{X}=(X,Y)^\intercal$ ühistihedus

$$f(x,y)=\begin{cases} cx^2, \text{ kui } 0<x<y<1,\\ 0, \text{ mujal .}\end{cases}$$

Leidke konstant c ning keskmised regressioonisõltuvused $y=g_1(x)$ ja $x=g_2(y)$.

Ülesanne 4.4. Olgu juhusliku vektori $\mathbf{X}=(X,Y)^\intercal$ ühistihedus

$$f(x,y)=\begin{cases} c, \text{ kui } 0\le x\le 1, \quad x^2\le y\le x,\\ 0, \text{ mujal.}\end{cases}$$

Leidke konstant c ning keskmised regressioonisõltuvused $y=g_1(x)$ ja $x=g_2(y)$.

Ülesanne 4.5. Olgu juhusliku vektori $\mathbf{X}=(X,Y)^\intercal$ ühistihedus

$$f(x,y)=\begin{cases} x\exp(-x(1+y)), \text{ kui } 0>x, \quad y>0,\\ 0, \text{ mujal.}\end{cases}$$

Leidke keskmised regressioonisõltuvused $y=g_1(x)$ ja $x=g_2(y)$.

Ülesanne 4.6. Olgu juhusliku vektori $\mathbf{X}=(X,Y)^\intercal$ ühistihedus

$$f(x,y)=\begin{cases} 24xy, \text{ kui } x+y\le 1, x\ge 0, y\ge 0,\\ 0, \text{ mujal .}\end{cases}$$

Leidke parim lineaarne ennustaja $L(X)=\beta_0+\beta_1 X$. Milline on tema R-ruut determinatsioonikordaja?

Ülesanne 4.7. Allugu lambipirni eluiga T eksponentjaotusele parameetriga λ. Olgu meil 15% lambipirnidest keskmise elueaga 9 kuud, 60% lambipirnidest 11 kuud ja 25% lambipirnidest 14 kuud. Me võtame suvalise lambipirni. Kui suure tõenäosusega peab see vastu üle aasta?

Ülesanne 4.8. Allugu projektori eluiga T eksponentjaotusele. Olgu selle projektori keskmine eluiga 10 tundi. Allugu loengute arv nädalas Poissoni jaotusele. Olgu nädalas keskmiselt 12 loengut ja eeldame, et igas loengus kasutatakse projektorit täpselt 1 tund. Leidke tõenäosus, et projektor töötab vähemalt nädala.

Ülesanne 4.9. Olgu aparaadi eluiga (aastates) T eksponentjaotusega parameetriga $m > 0$. Olgu parameeter m juhusliku suuruse M mingi väärtus. Selle juhusliku suuruse M tihedusfunktsioon

$$g(m) = \begin{cases} \dfrac{1}{m}, \text{ kui } m \in [1; e], \\ 0, \text{ mujal .} \end{cases}$$

Leidke

1) aparaadi tööea tihedusfunktsioon $f(t)$, kasutades tinglikustamise võtet;

2) aparaadi keskmine tööiga $E(T)$.

Ülesanne 4.10. Olgu juhuslikud suurused $N, X_1, X_2, ...$ sõltumatud. Olgu N Poissoni jaotusega parameetriga λ ja X_k, $k \geq 1$ Bernoulli jaotusega parameetriga $\dfrac{1}{2}$. Olgu meil juhuslikud suurused

$$Y_1 = \sum_{i=1}^{N} X_i$$

ja

$$Y_2 = N - Y_1$$

($Y_1 = 0$, kui $N = 0$). Tõestage, et juhuslikud suurused Y_1 ja Y_2 on sõltumatud ning leidke nende jaotused.

Ülesanne 4.11. Olgu juhuslik suurus Y binoomjaotusega $B(4, X)$, kus X allub ühtlasele jaotusele lõigus $[0, 1]$. Leidke tõenäosus $P(X \leq 0.5)$, kui on teada, et 1) $Y = 3$ ning 2) $Y = 1$.

Ülesanne 4.12. Olgu juhuslik suurus Y binoomjaotusega $B(n, X)$, kus X allub ühtlasele jaotusele lõigus $[0, 1]$. Leidke $E(Y)$, $D(Y)$ ning korrelatsioonikordaja $\text{corr}(X, Y)$.

Ülesanne 4.13. Olgu meil juhuslik suurus Y binoomjaotusega $B(5, X)$, kus juhusliku suuruse X tihedusfunktsioon

$$f(x) = \begin{cases} 0, \text{ kui } x \notin [0;1], \\ 2x, \text{ kui } x \in [0;1]. \end{cases}$$

1) Leidke tõenäosus $P(X < 0.5)$.

2) Milliseks aga kujuneb punktis 1 toodud tõenäosus juhul, kui on teada, et $Y = 3$ ehk teadmise juures, et 5 sõltumatust katsest õnnestus 3?

Ülesanne 4.14. Allugu juhuslik suurus Y ühtlasele jaotusele vahemikus $(-X; X)$, kus juhuslik suurus X on ühtlase jaotusega vahemikus $(1; 2)$. Leidke tihedusfunktsioon $f(y)$ ning keskväärtus $E(Y)$.

Ülesanne 4.15. Viimaste aastate uuringud on näidanud, et 5% naistest põeb rinnavähki. Rinnavähki haigestumist saab kontrollida mammograafiga. Olgu selle testi võimsus 0.9. Kui suur on tõenäosus, et mammograafi näidu põhjal positiivse vastuse saanud naine põeb rinnavähki, kui testi olulisuse nivoo $\beta = 0.05$?

Ülesanne 4.16. Laual on kolm münti, millest 2 on sümmeetrilised, kolmas münt on aga eriline. Sellel erilisel mündil on vapi tuleku tõenäosus 0.55. Me võtame laualt suvalise mündi ja saame seda visates vapi. Kuidas mõjutas vapi tulek sündmuse $A =$ „münt on eriline“ tõenäosust võrreldes tavamündi saamise tõenäosusega?

Ülesanne 4.17. Müügikõlblike ravimite puhul on nõuetele mittevastava medikamendi leidmise tõenäosus 0.01. Müügikõlbmatute puhul aga on see tõenäosus 0.05. Hiirte ja rottide peal tehtud katsete põhjal on müügikõlblikke ravimeid 20%. Testiti sõltumatult 100 ravimit. Minimaalselt mitme nõuetele mittevastava ravimi leidmise puhul ollakse sunnitud partiid mitte müügile laskma? Seda eeldusel, et trahvisumma müügikõlbmatu ravimi turustamise eest on 10 korda suurem kui saamata jäänud kasum müügikõlbliku ravimi partii välja praakimise puhul.

Ülesanne 4.18. Allugu juhuslik suurus X normaaljaotusele $\mathcal{N}(\mu, \sigma)$. Koostage hüpoteeside paarile

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2, \\ H_1 : \sigma^2 = \sigma_1^2 \end{cases}$$

vastav Neyman-Pearsoni kriteerium.

Ülesanne 4.19. Allugu patarei eluiga aastates eksponentjaotusele, mille parameetri M jaotust kirjeldab tihedusfunktsioon

$$f(m) = \begin{cases} \dfrac{m}{2}, \text{ kui } m \in (0;2), \\ 0, \text{ mujal .} \end{cases}$$

Mõõdeti 5 patarei vastupidavust aastates. Tulemusteks saadi (0.9, 1.2, 1.6, 0.8, 1.1). Kui suur on pärast selliseid mõõtmistulemusi tõenäosus, et patarei peab keskmiselt vastu üle aasta? Võrrelge seda tõenäosust ülesandes 1.14 leitud tõenäosusega.

Ülesanne 4.20. Allugu mingis regioonis tulekahjude hulk aastas Poissoni jaotusele. Olgu Poissoni jaotuse parameeter $M \sim \mathcal{E}(1)$. Leidke selle regiooni keskmine tulekahjude hulk aastas, kui on teada, et eelmisel aastal toimus seal 3 tuleõnnetust.

Ülesanne 4.21. Olgu meil juhuslik suurus $X \sim B(N, p)$, kus sõltumatute katsete hulk N allub Poissoni jaotusele parameetriga 2. Olgu katse õnnestumise tõenäosus $p = 0.7$. Kui suure tõenäosusega oli sõltumatute katsete hulk 3, kui on teada, et katse õnnestus 1 korral.

Ülesanne 4.22. Kirjeldagu seisunditest E_1, E_2, E_3 koosnevat Markovi ahelat üleminekumaatriks

$$P = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0.1 & 0.8 & 0.1 \\ 0.5 & 0.1 & 0.4 \end{pmatrix}.$$

Leidke antud 3 seisundi statsionaarne jaotus π^*.

Ülesanne 4.23. Analüüsige Markovi ahela seisundite statsionaarset jaotust järgmiste üleminekumaatriksite korral:

$$\mathbf{P}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{P}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ ja } \mathbf{P}_3 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}.$$

Ülesanne 4.24. Leidke näites 4.21 toodud elektriskeemile tõenäosus, et see peab vastu üle aasta, kui seisunditele E_1, E_2 ning E_3 vastavad keskmised eluead on 10 kuud, 13 kuud ja 18 kuud. Aluseks võtta seisundite statsionaarne jaotus.

Ülesanne 4.25. Leidke näites 4.25 toodud Markovi ahela üleminekumaatriks $\mathbf{P}$, kui

$$\pi_{00} = 0.3,\ \pi_{01} = 0.2,\ \pi_{10} = 0.4 \text{ ja } \pi_{11} = 0.1.$$

Ülesanne 4.26. Allugu uuritav suurus X_i, $i = 1, 2, ..., n$ normaaljaotusele $\mathcal{N}(\mu, \sigma)$. Olgu parameetri μ eeljaotuseks normaaljaotus $\mathcal{N}(0, \tau)$ ning parameetri σ^2 eeljaotuseks eksponentjaotus $\mathcal{E}(1)$. Leidke Gibbsi valikut rakendades järeljaotus $\pi(\mu, \sigma^2)$.

Ülesanne 4.27. Allugu juhuslik suurus X eksponentjaotusele $\mathcal{E}(\Lambda)$. Juhuslik suurus Λ allugu Reighleigh' jaotusele, mille parameeter $h = 1$. Koostada näite 4.26 eeskujul eksponentjaotuse parameetri λ muutlikkusele Metropolis-Hastingsi algoritm eeldusel, et muut $\Delta\lambda$ allub normaaljaotusele keskväärtusega 0 ning standardhälbega 0.2.

Ülesanne 4.28. Realiseerige näite 4.27 algoritm juhul, kui üleminekutuumaks on normaaljaotusele $\mathcal{N}(-2, 2)$ vastav tõenäosus. Milliseid erinevusi märkate võrreldes näitega 4.27.

Ülesanne 4.29. Allugu patarei eluiga T (aastates) eksponentjaotusele, mille parameeter M allugu gammajaotusele $G(0.1, 0.1)$. Koostage parameetri M järeljaotuse mudel ülesande 4.19 andmete põhjal. Modelleerimiseks võiks kasutada tarkvara OpenBugs või Stan.

Lisad

Lisa 1. Laplace'i veafunktsiooni $\Phi(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ väärtused

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.000	0.004	0.008	0.012	0.016	0.020	0.024	0.028	0.032	0.036
0.1	0.040	0.044	0.048	0.052	0.056	0.060	0.064	0.067	0.071	0.075
0.2	0.079	0.083	0.087	0.091	0.095	0.099	0.103	0.106	0.110	0.114
0.3	0.118	0.122	0.126	0.129	0.133	0.137	0.141	0.144	0.148	0.152
0.4	0.155	0.159	0.163	0.166	0.170	0.174	0.177	0.181	0.184	0.188
0.5	0.191	0.195	0.198	0.202	0.205	0.209	0.212	0.216	0.219	0.222
0.6	0.226	0.229	0.232	0.236	0.239	0.242	0.245	0.249	0.252	0.255
0.7	0.258	0.261	0.264	0.267	0.270	0.273	0.276	0.279	0.282	0.285
0.8	0.288	0.291	0.294	0.297	0.300	0.302	0.305	0.308	0.311	0.313
0.9	0.316	0.319	0.321	0.324	0.326	0.329	0.331	0.334	0.336	0.339
1.0	0.341	0.344	0.346	0.348	0.351	0.353	0.355	0.358	0.360	0.362
1.1	0.364	0.367	0.369	0.371	0.373	0.375	0.377	0.379	0.381	0.383
1.2	0.385	0.387	0.389	0.391	0.393	0.394	0.396	0.398	0.400	0.401
1.3	0.403	0.405	0.407	0.408	0.410	0.411	0.413	0.415	0.416	0.418
1.4	0.419	0.421	0.422	0.424	0.425	0.426	0.428	0.429	0.431	0.432
1.5	0.433	0.434	0.436	0.437	0.438	0.439	0.441	0.442	0.443	0.444
1.6	0.445	0.446	0.447	0.448	0.449	0.451	0.452	0.453	0.454	0.454
1.7	0.455	0.456	0.457	0.458	0.459	0.460	0.461	0.462	0.462	0.463
1.8	0.464	0.465	0.466	0.466	0.467	0.468	0.469	0.469	0.470	0.471
1.9	0.471	0.472	0.473	0.473	0.474	0.474	0.475	0.476	0.476	0.477
2.0	0.477	0.478	0.478	0.479	0.479	0.480	0.480	0.481	0.481	0.482
2.1	0.482	0.483	0.483	0.483	0.484	0.484	0.485	0.485	0.485	0.486
2.2	0.486	0.486	0.487	0.487	0.487	0.488	0.488	0.488	0.489	0.489
2.3	0.489	0.490	0.490	0.490	0.490	0.491	0.491	0.491	0.491	0.492
2.4	0.492	0.492	0.492	0.492	0.493	0.493	0.493	0.493	0.493	0.494
2.5	0.494	0.494	0.494	0.494	0.494	0.495	0.495	0.495	0.495	0.495
2.6	0.495	0.495	0.496	0.496	0.496	0.496	0.496	0.496	0.496	0.496
2.7	0.497	0.497	0.497	0.497	0.497	0.497	0.497	0.497	0.497	0.497
2.8	0.497	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.498
2.9	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.499	0.499	0.499
3.0	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499

Lisa 2. Studenti t-testi kriitilised väärtused erinevate olulisuse nivoode β ning vabadusastmete arvu k korral

β **kahepoolne** / k	0.01	0.02	0.04	0.05	0.1	0.2
2	9.92	6.96	4.85	4.30	2.92	1.89
3	5.84	4.54	3.48	3.18	2.35	1.64
4	4.60	3.75	3.00	2.78	2.13	1.53
5	4.03	3.36	2.76	2.57	2.02	1.48
6	3.71	3.14	2.61	2.45	1.94	1.44
7	3.50	3.00	2.52	2.36	1.89	1.41
8	3.36	2.90	2.45	2.31	1.86	1.40
9	3.25	2.82	2.40	2.26	1.83	1.38
10	3.17	2.76	2.36	2.23	1.81	1.37
11	3.11	2.72	2.33	2.20	1.80	1.36
12	3.05	2.68	2.30	2.18	1.78	1.36
13	3.01	2.65	2.28	2.16	1.77	1.35
14	2.98	2.62	2.26	2.14	1.76	1.35
15	2.95	2.60	2.25	2.13	1.75	1.34
16	2.92	2.58	2.24	2.12	1.75	1.34
17	2.90	2.57	2.22	2.11	1.74	1.33
18	2.88	2.55	2.21	2.10	1.73	1.33
19	2.86	2.54	2.20	2.09	1.73	1.33
20	2.85	2.53	2.20	2.09	1.72	1.33
21	2.83	2.52	2.19	2.08	1.72	1.32
22	2.82	2.51	2.18	2.07	1.72	1.32
23	2.81	2.50	2.18	2.07	1.71	1.32
24	2.80	2.49	2.17	2.06	1.71	1.32
25	2.79	2.49	2.17	2.06	1.71	1.32
26	2.78	2.48	2.16	2.06	1.71	1.31
27	2.77	2.47	2.16	2.05	1.70	1.31
28	2.76	2.47	2.15	2.05	1.70	1.31
29	2.76	2.46	2.15	2.05	1.70	1.31
30	2.75	2.46	2.15	2.04	1.70	1.31
50	2.68	2.40	2.11	2.01	1.68	1.30
80	2.64	2.37	2.09	1.99	1.66	1.29
120	2.62	2.36	2.08	1.98	1.66	1.29
β **ühepoolne** / k	0.005	0.01	0.02	0.025	0.05	0.1

Lisa 3. χ^2-jaotuse α-täiendkvantiilid erinevate vabadusastmete arvu k korral

k \ α	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
1	6.63	5.02	3.84	2.71	0.02	0.00	0.00	0.00
2	9.21	7.38	5.99	4.61	0.21	0.10	0.05	0.02
3	11.34	9.35	7.81	6.25	0.58	0.35	0.22	0.11
4	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.30
5	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55
6	16.81	14.45	12.59	10.64	2.20	1.64	1.24	0.87
7	18.48	16.01	14.07	12.02	2.83	2.17	1.69	1.24
8	20.09	17.53	15.51	13.36	3.49	2.73	2.18	1.65
9	21.67	19.02	16.92	14.68	4.17	3.33	2.70	2.09
10	23.21	20.48	18.31	15.99	4.87	3.94	3.25	2.56
11	24.73	21.92	19.68	17.28	5.58	4.57	3.82	3.05
12	26.22	23.34	21.03	18.55	6.30	5.23	4.40	3.57
13	27.69	24.74	22.36	19.81	7.04	5.89	5.01	4.11
14	29.14	26.12	23.68	21.06	7.79	6.57	5.63	4.66
15	30.58	27.49	25.00	22.31	8.55	7.26	6.26	5.23
16	32.00	28.85	26.30	23.54	9.31	7.96	6.91	5.81
17	33.41	30.19	27.59	24.77	10.09	8.67	7.56	6.41
18	34.81	31.53	28.87	25.99	10.86	9.39	8.23	7.01
19	36.19	32.85	30.14	27.20	11.65	10.12	8.91	7.63
20	37.57	34.17	31.41	28.41	12.44	10.85	9.59	8.26
21	38.93	35.48	32.67	29.62	13.24	11.59	10.28	8.90
22	40.29	36.78	33.92	30.81	14.04	12.34	10.98	9.54
23	41.64	38.08	35.17	32.01	14.85	13.09	11.69	10.20
24	42.98	39.36	36.42	33.20	15.66	13.85	12.40	10.86
25	44.31	40.65	37.65	34.38	16.47	14.61	13.12	11.52
26	45.64	41.92	38.89	35.56	17.29	15.38	13.84	12.20
27	46.96	43.19	40.11	36.74	18.11	16.15	14.57	12.88
28	48.28	44.46	41.34	37.92	18.94	16.93	15.31	13.56
29	49.59	45.72	42.56	39.09	19.77	17.71	16.05	14.26
30	50.89	46.98	43.77	40.26	20.60	18.49	16.79	14.95
35	57.34	53.20	49.80	46.06	24.80	22.47	20.57	18.51
40	63.69	59.34	55.76	51.81	29.05	26.51	24.43	22.16
60	88.38	83.30	79.08	74.40	46.46	43.19	40.48	37.48
80	112.33	106.63	101.88	96.58	64.28	60.39	57.15	53.54
100	135.81	129.56	124.34	118.50	82.36	77.93	74.22	70.06

Lisa 4. Fisheri F-jaotuse 0.95-kvantiilid erinevate vabadusastmete arvude n ja m korral

n \ m	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	19.0	9.6	6.9	5.8	5.1	4.7	4.5	4.3	4.1	4.0	3.9	3.8	3.7	3.7
3	19.2	9.3	6.6	5.4	4.8	4.3	4.1	3.9	3.7	3.6	3.5	3.4	3.3	3.3
4	19.2	9.1	6.4	5.2	4.5	4.1	3.8	3.6	3.5	3.4	3.3	3.2	3.1	3.1
5	19.3	9.0	6.3	5.1	4.4	4.0	3.7	3.5	3.3	3.2	3.1	3.0	3.0	2.9
6	19.3	8.9	6.2	5.0	4.3	3.9	3.6	3.4	3.2	3.1	3.0	2.9	2.8	2.8
7	19.4	8.9	6.1	4.9	4.2	3.8	3.5	3.3	3.1	3.0	2.9	2.8	2.8	2.7
8	19.4	8.8	6.0	4.8	4.1	3.7	3.4	3.2	3.1	2.9	2.8	2.8	2.7	2.6
9	19.4	8.8	6.0	4.8	4.1	3.7	3.4	3.2	3.0	2.9	2.8	2.7	2.6	2.6
10	19.4	8.8	6.0	4.7	4.1	3.6	3.3	3.1	3.0	2.9	2.8	2.7	2.6	2.5
11	19.4	8.8	5.9	4.7	4.0	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5
12	19.4	8.7	5.9	4.7	4.0	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.5	2.5
13	19.4	8.7	5.9	4.7	4.0	3.6	3.3	3.0	2.9	2.8	2.7	2.6	2.5	2.4
14	19.4	8.7	5.9	4.6	4.0	3.5	3.2	3.0	2.9	2.7	2.6	2.6	2.5	2.4
15	19.4	8.7	5.9	4.6	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.5	2.4
16	19.4	8.7	5.8	4.6	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.4	2.4
17	19.4	8.7	5.8	4.6	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.4	2.4
18	19.4	8.7	5.8	4.6	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.4	2.4
19	19.4	8.7	5.8	4.6	3.9	3.5	3.2	2.9	2.8	2.7	2.6	2.5	2.4	2.3
20	19.4	8.7	5.8	4.6	3.9	3.4	3.2	2.9	2.8	2.6	2.5	2.5	2.4	2.3
21	19.4	8.7	5.8	4.5	3.9	3.4	3.1	2.9	2.8	2.6	2.5	2.4	2.4	2.3
22	19.5	8.6	5.8	4.5	3.9	3.4	3.1	2.9	2.8	2.6	2.5	2.4	2.4	2.3
23	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3
24	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
25	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
26	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
27	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
28	19.5	8.6	5.8	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
29	19.5	8.6	5.7	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.3
30	19.5	8.6	5.7	4.5	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.4	2.3	2.2

Ülesannete vastused ja lahendused

Peatükk 1

1.1 Hinnang $\overline{x}_1$ on nihketa ja mõjus ning hinnang $\overline{x}_2$ on nihkega ja ei ole mõjus; **1.2** $\mathrm{MSE}(\widehat{\theta}_1) = 10$ ning $\mathrm{MSE}(\widehat{\theta}_2) = 4 + \dfrac{\theta^2}{4}$; **1.4** Normaaljaotus $\mathcal{N}(k, \dfrac{\sqrt{0.12}}{3})$; **1.5** $p^* = \dfrac{1}{\overline{x}}$; **1.6** $a^* = \overline{x}$; **1.7** $I_{0.95} \approx \left[\dfrac{1}{\overline{x} + 1.96\frac{\overline{x}}{6}}; \dfrac{1}{\overline{x} - 1.96\frac{\overline{x}}{6}}\right]$ **1.8** $I_{0.95} \approx [0.023; 0.355]$; **1.9** $I_{0.95} \approx [3.8; 4.9]$; **1.10** $I_{0.95} \approx [6.72; 13.28]$, suhteline viga on üle 10%; **1.11** 27 mõõtmist; **1.12** Minimaalne küsitletute hulk $n = 381$; **1.13** $\widehat{p} = 0.5$ korral; **1.14** $m^* \approx 0.893$, $P(\text{„elab üle 1 aasta“}) \approx 0.41$; **1.15** 1) Studenti t-test, võib lugeda tõestatuks hüpoteesi, et seina paksus on üle 4 mm, sest $p\text{-}value < 0.001$ ning 2) keskväärtuse $I_{0.95} \approx [4.019; 4.097]$ ja standardhälbe $I_{0.95} \approx [0.063; 0.113]$; **1.16** $\overline{x} = 201.2$ m, $I_{0.95} = [196.0; 206.4]$ ning $s^2 = 37.51$ m^2, $I_{0.95} = [17.74; 125.02]$; **1.17** Inseneril A $h(\mu) = 1 + \Phi\left(\dfrac{9.9 - \mu}{0.3}\sqrt{8}\right) - \Phi\left(\dfrac{10.1 - \mu}{0.3}\sqrt{8}\right)$ ja $h(10.2) \approx 0.83$ ning inseneril B $h(\mu) = 1 + \Phi\left(\dfrac{9.8 - \mu}{0.3}\sqrt{12}\right) - \Phi\left(\dfrac{10.2 - \mu}{0.3}\sqrt{12}\right)$ ja $h(10.2) \approx 0.5$; **1.18** Poissoni jaotusega, $\lambda^* = 2.1$, $h \approx 3.2$, $p\text{-}value \approx 0.52$, Poissoni jaotuse eeldus jääb kehtima; **1.19** Võib nullhüpoteesi kummutada, sest $p\text{-}value < 0.001$; **1.20** 2) Studenti t-testi erinevate valimite puhul, 3) mõlema rahvusgrupi IQ on normaaljaotusega ning 4) ei või, sest $p\text{-}value \approx 0.359$; **1.21** Võib väita, sest $p\text{-}value \approx 0.004$; **1.22** 1) $p\text{-}value \approx 0.009$, 2) $p\text{-}value \approx 0.063$ ning 3) $p\text{-}value \approx 0.164$; **1.23** 1) $p\text{-}value \approx 0.176$, 2) $p\text{-}value \approx 0.015$ ning 3) $p\text{-}value \approx 0.0013$; **1.24** Võib väita, sest $p\text{-}value \approx 0.015$; **1.25** 1) $p\text{-}value \approx 0.045$, 2) $\gamma_2 \approx 0.093$ ning 3) $h(13.5) \approx 0.907$; **1.26** 1) Ei või väita, sest $p\text{-}value \approx 0.57$, 2) vähemalt 245 viset; **1.27** Ei paranenud oluliselt, sest $p\text{-}value \approx 0.28$; **1.28** Ei või ümber lükata, Fisheri F-testile vastav $p\text{-}value \approx 0.083$

Peatükk 2

2.1 $\mathrm{corr}(X_1, X_2) \approx 0.868$; **2.2** $\beta_0 \approx 2.63$ ning $\beta_1 \approx 2.78$ **2.3** $R^2 \approx 0.674$; **2.4** 1) ≈ 0.407 $\mathrm{M}^{-1}\mathrm{cm}^{-1}$ võrra, 2) ≈ 0.255 $\mathrm{M}^{-1}\mathrm{cm}^{-1}$ ning 3) $R^2 \approx 0.99$; **2.5** $\mathbf{r} \approx 0.66$; **2.6** 1) $Y = -14.1 + 1.77X_1 + 2.35X_2 + 0.41X_3 + 0.75X_4$, 2) *p-value* $= 2.61 \cdot 10^{-6}$ 3) olulised on tunnused X_1 ja X_2 ning 4) $R^2 \approx 0.98$; **2.7** Sõltub oluliselt aastaajast, *p-value* < 0.001; **2.8** Ei või ümber lükata, *p-value* ≈ 0.995; **2.9** Testi tulemust mõjutas oluliselt nii sugu (*p-value* < 0.001) kui ka proteiini tarbimine (*p-value* ≈ 0.017,) koosmõju ei olnud oluline (*p-value* ≈ 0.226); **2.10** $z_i = \ln(\lambda_i) + \dfrac{y_i - \lambda_i}{\lambda_i}$, diagonaalmaatriksi $\mathbf{W}$ peadiagonaali element $w_i = \lambda_i$; **2.11** Ei või kummutada, teststatistiku väärtus on 2.7, *p-value* ≈ 0.1; **2.12** Võib kummutada, teststatistiku väärtus on 6.25, *p-value* ≈ 0.012; **2.13** $D(\mathbf{y}, \widehat{\nu}) = 2\sum_{i=1}^{n}\left\{\frac{\mathrm{emp}_i - \mathrm{teor}_i}{\mathrm{teor}_i} - \ln\left(\frac{\mathrm{emp}_i}{\mathrm{teor}_i}\right)\right\}$; **2.14** $\theta = \ln(1-p)$ ning $p = \exp(\theta) - 1$; **2.15** 2) $\theta = \ln\left(\dfrac{a}{1+a}\right)$ ja $b(\theta) = \ln\left(\dfrac{1}{1+\exp(\theta}\right)$, 3) $E(Y) = b'(\theta) = a$ ning 4) $a = \dfrac{\exp(\theta)}{1-\exp(\theta)}$; **2.16** 1) $p = \dfrac{\exp(-2.49 + 0.95 \cdot \mathrm{Tunnid})}{1 + \exp(-2.49 + 0.95 \cdot \mathrm{Tunnid})}$, 2) Ei sõltunud oluliselt, *p-value* > 0.1 ning 3) Skoori testi põhjal leitud olulisustõenäosus *p-value* ≈ 0.56, tuleb jääda hüpoteesi $p = 0.5$ juurde; **2.17** Vähim AIC väärtus vastab mudelile, kuhu on kaasatud nii faktorid X_1, X_2 kui ka X_3; **2.18** 2) Kommunaliteedid on järgmised: AEG – 0.680, PINDALA – 0.774, MAKSUMUS – 0.814, ARVUKUS – 0.770, PIKKUS – 0.874, SPAN – 0.787 ja KEERUKUS – 0.905 ning 3) $\approx 80\%$; **2.19** Maatriks $\mathbf{R}$ ei ole positiivselt määratud; **2.20** **A** ei või olla, **B** võib olla, **C** ei või olla ning **D** võib olla; **2.21** Esimene peakomponent kirjeldab $\approx 59.3\%$ koguhajuvust, teine peakomponent $\approx 24\%$

Peatükk 3

3.1 $f_{T_{(1)}} = \dfrac{4}{5}\exp\left(-\dfrac{4x}{5}\right)$ ning $f_{T_{(8)}} = \dfrac{4}{5}\left(1-\exp\left(-\dfrac{x}{10}\right)\right)^7 \exp\left(-\dfrac{x}{10}\right)$; **3.2** $P(X_{(1)} > 0.1) \approx 0.857$ ning $P(X_{(3)} \leq 0.9) \approx 0.0911$; **3.3** $F_{X_{(1)}} = 1 - \left(1 - \dfrac{\sin(x)+1}{2}\right)^6$ ning $f_{X_{(1)}} = 6\left(1 - \dfrac{\sin(x)+1}{2}\right)^5 \dfrac{\cos(x)}{2}$ ja

$F_{X_{(6)}} = \left(\frac{\sin(x)+1}{2}\right)^6$ ning $f_{X_{(6)}} = 6\left(\frac{\sin(x)+1}{2}\right)^5 \frac{\cos(x)}{2}$;
3.4 $F_{X_{(5)}}(x) = \sum_{k=5}^{9} \frac{9!}{k!(9-k)!} \frac{x^{2k}}{4^k} \left(1-\frac{x^2}{4}\right)^{9-k}$, $x \in [0;2]$ ning
$f_{X_{(5)}}(x) = \frac{x^8}{967680}\left(1-\frac{x^2}{4}\right)^8 \frac{x}{2}$, kui $x \in [0;2]$ **3.5** $G(x) = \exp(-x^{-2})$ ehk Fréchet' jaotus; **3.6** $E(X_{(2)}) = \frac{7}{24}$; **3.7** $\approx 6.05 \cdot 10^{-3}$; **3.8** $\approx 9.57 \cdot 10^{-5}$; **3.9** 1) $\rho(X,Y) = -1$, 2) $\rho(X,Y) = 1, x \neq 0$, 3) $\rho(X,Y) = 0$ ning 4) $\rho(X,Y) = -1$; **3.10** $\mathrm{corr}(X,Y) \approx 0.9$, $\rho(X,Y) = 0.6$ ning $\tau(X,Y) = 0.4$; **3.11** $\rho(X,Y) = 0.9$ ning $\tau(X,Y) = 0.8$; **3.12** Studenti t-jaotuse abil $I_{0.95} \approx [1.35, 1.74]$, meetodil *jacknife* leitud $I_{0.95} \approx [1.14, 1.95]$; **3.13** Ei ole piisav, sest *p-value* > 0.05; **3.14** $I_{0.95} \approx [1.65; 2.25]$; **3.15** $I_{0.95} \approx [48.4; 61.6]$ t-jaotuse põhjal ning $I_{0.95} \approx [51.1; 59.7]$ $\mathrm{BC_a}$ meetodil; **3.16** Pärast nihkeparandit $\alpha_1 \approx 0.27$ ning $\alpha_2 \approx 0.99$; **3.17** Ei või ümber lükata, *p-value* ≈ 0.61; **3.18** Tuleb jääda nullhüpoteesi juurde, *p-value* ≈ 0.55; **3.19** Ei saa tõestada, *p-value* ≈ 0.34; **3.20** Eksponentjaotuse eelduse võib ümber lükata, $\sqrt{n}D \approx 2.46$ ning *p-value* < 0.005; **3.23** $\alpha_1 \approx 0.09$ ning $\alpha_2 \approx 0.998$; **3.24** Näpunäide: kasutage permutatsiooni lemmat; **3.25** Ei või ümber lükata, *p-value* ≈ 0.26

Peatükk 4

4.1 $\gamma(X,Y) \approx 0.232$, $\gamma(Y,X) \approx 0.202$ ning $\chi(X,Y) \approx 0.23$;
4.2 $\mathrm{corr}(X,Y) \approx 0.079$, $\mathrm{corr}(X,Z) \approx 0.079$ ning $\mathrm{corr}(Y,Z) = 1$; $\gamma(X,Y) = 0.079$ ning $\gamma(Y,X) = 1$; $\gamma(X,Z) \approx 0.079$ ning $\gamma(Z,X) = 1$; $\gamma(Y,Z) = 1$ ning $\gamma(Z,Y) = 1$; **4.3** $c = 12$, $g_1(x) = \frac{1+x}{2}$, kui $x \in (0;1)$ ning $g_2(y) = \frac{3}{4}y$, kui $y \in (0;1)$; **4.4** $c = 6$, $g_1(x) = \frac{x+x^2}{2}$ ning $g_2(y) = \frac{y+\sqrt{y}}{2}$; **4.5** $g_1(x) = \frac{1}{x}$ ning $g_2(y) = \frac{2}{1+y}$; **4.6** $L(X) = \frac{2}{5} - \frac{2}{3}(X - \frac{2}{5})$, $R^2 = \frac{4}{9}$; **4.7** $P(T > 1 \text{ aasta}) \approx 0.347$; **4.8** $P(T > 1 \text{ nädal}) = \exp\left(12(e^{-\frac{1}{10}} - 1)\right) \approx 0.319$; **4.9** 1) $f(t) = \frac{\exp(-t) - \exp(-et)}{t}$ ning 2) $E(T) = 1 - \frac{1}{e}$; **4.10** 1)

$Y_1 \sim Po(\frac{\lambda}{2})$ ning $Y_2 \sim Po(\frac{\lambda}{2})$; **4.11** 1) $P(X \leq 0.5 \mid Y = 3) = 0.1875$ ning 2) $P(X \leq 0.5 \mid Y = 1) = 0.8125$; **4.12** $E(Y) = 2$, $D(Y) = 2$ ning $\text{corr}(X, Y) = \frac{\sqrt{6}}{3}$; **4.13** 1) $P(X < 0.5) = 0.25$ ning 2) $P(X < 0.5 \mid Y = 3) \approx 0.227$; **4.14** Ühtlase jaotusega vahemikus $(0; \ln(\sqrt{2}))$, $E(Y) = \frac{\ln(2)}{4}$; **4.15** ≈ 0.486; **4.16** Suurenes 1.1 korda; **4.17** Miinimaalselt 2 ravimi puhul; **4.18** $\text{LR} = \frac{\sigma_0}{\sigma_1} \exp\left(\frac{(x-\mu)^2(\sigma_0^2 - \sigma_1^2)}{2\sigma_0^2\sigma_1^2}\right)$; **4.19** $P(\text{„elab üle 1 aasta“}) \approx 0.355$; **4.20** Keskmine tulekahjude hulk $E(M) = 2$; **4.21** $P(N = 3 \mid X = 1) = 0.06\exp(-0.6)$; **4.22** $\pi^* = \begin{pmatrix} 0.2 & 0.6 & 0.2 \end{pmatrix}$; **4.23** $\mathbf{P}_1$ – kõik jaotused on statsionaarsed, $\mathbf{P}_2$ puhul $\pi_1 = \pi_2 = 0.5$ ning $\mathbf{P}_3$ – $\pi^* = \begin{pmatrix} 0 & 1 \end{pmatrix}$; **4.24** $P(\text{„skeem peab vastu üle aasta“}) \approx 0.403$;

4.25 $\mathbf{P} = \begin{pmatrix} 0.257 & 0.171 & 0.457 & 0.114 \\ 0.4 & 0.267 & 0.267 & 0.067 \\ 0.257 & 0.171 & 0.457 & 0.114 \\ 0.4 & 0.267 & 0.267 & 0.067 \end{pmatrix}$;

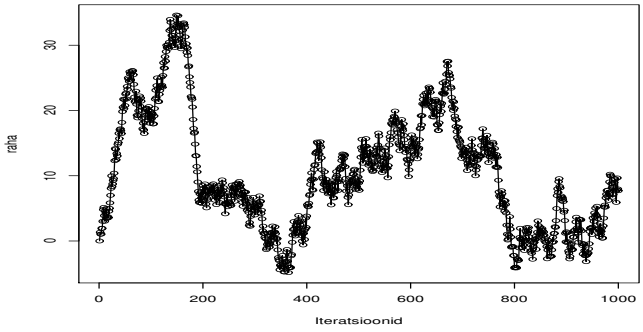
4.26 $\pi(\mu, \sigma^2) \propto \exp\left(-\frac{n\tau^2(\overline{x} - \mu)^2 + \sigma^2(2\tau\sigma^2 + \mu^2)}{2\sigma^2\tau^2}\right)$;

4.27 Vastuvõtmise kriteerium

$$\alpha(\lambda, \lambda + \Delta\lambda) = \min\left\{1, \frac{(\lambda + \Delta\lambda)\exp\{-(\lambda + \Delta\lambda)(\lambda + \Delta\lambda + 1)\}}{\lambda\exp\{-\lambda(\lambda + 1)\}}\right\},$$

kui $\Delta\lambda \sim \mathcal{N}(0, 0.2)$;

4.28 Muutlikkust kirjeldav graafik võib olla järgmine:



Kirjandus

[1] Anderson, T. W., Darling, D. A. (1954) A Test of Goodness-of-Fit. *Journal of the American Statistical Association*, **49**, lk 765–769.

[2] Arumägi, E., Kalamees, T., Pihlak, M. (2015) Reliability of unterior thermal isulation as a retrofit measure in historic wooden apartment buildings in cold climite. *Energy Procedia*, **78**, lk 871–876.

[3] Blom, G. (1989) *Probabilty and Statistics*: *Theory and Applications*, Springer Verlag.

[4] Carlin, B. P., Gelfand, A. E, Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, **41**, lk 389– 405.

[5] Carpenter, B. jt (2017) Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, **76**, nr 1.

[6] Corder, G. W., Foremann, D. I. (2014) *Nonparametric Statistics*, Wiley.

[7] Davison, A. C, Hinkley, D. V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press.

[8] Efron, B. (1979) Bootstrap methods: another look at the jacknife. *The Annals of Statistics*, **7**, lk 1–26.

[9] Efron, B., Tibshirani, Robert J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall.

[10] Efron, B. (2012) Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, **6**, nr 4, lk 1971–1997.

[11] Feller, W. (1948) On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions. *The Annals of Mathematical Statistics*, **19**, nr 2, lk 177–189.

[12] Gamerman, D., Lopes, Herbert F. (2006) *Markov Chain Monte Carlo*, Chapman & Hall.

[13] George, E. I., Makov, U. E. ja Smith, A. F. M. (1993) Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, **20**, lk 147– 156.

[14] Gnedenko, B. (1943) Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, **44**, lk 423–453.

[15] Gurski, J. (1986) *Tõenäosusteooria ja matemaatilise statistika elemente*, Tallinn Valgus.

[16] Gut, A. (2009) *An Intermediate Course in Probability*, Springer.

[17] Hastings W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, lk 97–109.

[18] Hoff, Peter D. (2009) *A First Course in Bayesian Statistical Methods*, Springer.

[19] Jayakumar, G. D. S. (2015) Exact distribution of Cook's distance and identification of influential observations. *Hacettepe J. Math. Stat.*, **44**, lk 165–178.

[20] Keres, K., Levin, A. (2006) *Matemaatilise statistika ülesannete kogu*, Tallinna Tehnikaülikooli kirjastus.

[21] Kiviste, A. (1999) *Matemaatiline statistika MS Excel keskkonnas*, GT Tarkvara OÜ. Tallinn.

[22] Kollo, T. (2004) *Monte Carlo meetodid*, Tartu Ülikooli Kirjastus.

[23] Kruschke, John K. (2015) *Doing Bayesian Data Analysis*, Elsevier.

[24] Lahiri, S. N. (2003) *Resampling methods for Dependent Data*, Springer-Verlag New York, Inc.

[25] Little, R. J. A. (1978). Generalized Linear Models for Cross-Classified Data from the WFS. *World Fertility Survey Technical Bulletins*, nr 5.

[26] Liu, P. ja Gene Hwang, J. T. (2007) Quick calculation for sample size while controlling false discovery rate with application to microarray data. textitBioinformatics, **23**(6), lk 739–746.

[27] Loone, L. ja Soomer V. (2009) *Matemaatilise analüüsi algkursus*, Tartu Ülikooli Kirjastus.

[28] Maivali, Ü. (2015) *Interpreting Biomedical Science*, Elsevier.

[29] Mendenhall, W., Sincich, T. (2007) *Statistics for engineering and the sciences*, Pearson Prentice Hall.

[30] Metropolis, N. jt (1953) Equations of step calculation by fast computing machine. *Journal of Chemical Physics*, **21**, lk 1087–1091.

[31] Montgomery, D. C., Runger, G. C., Huble. N. F. (2010) *Engineering Statistics*, John Wiley & Sons. Inc.

[32] Patterson, N., Price, A. L., Reich, D. (2006) Population Structure and Eigenanalysis. *PLOS Genetics*, http://dx.doi.org/10.1371/journal.pgen.0020190.

[33] Pihlak, M. (2005) Mängija laostumise probleem. *Eesti Matemaatika Selts*, Aastaraamat 2005, lk 60–64.

[34] Pärna, K. (2013) *Tõenäosusteooria algkursus*, Tartu Ülikooli Kirjastus.

[35] Puusemp, P. (2008) *Lineaaralgebra*, Avita.

[36] Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*, URL: http://data.princeton.edu/wws509/notes/

[37] Ryan, T. P. (2009) *Modern Regression Methods*, John Wile & Sons. Inc.

[38] Smirnov, N. V. (1948) Table for estimating goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, **19**, lk 279–281.

[39] Storey, John D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**(6), lk 2013–2035.

[40] Šapiro, S. S. ja Wilk, M. P. (1965) An analysis of variance test for normality (complete samples). *Biomatrika*, **52**(3-4). lk 591–611.

[41] Tammeraid, I. (2004) *Tõenäosusteooria ja matemaatiline statistika*, Tallinna Tehnikaülikooli kirjastus.

[42] Tiit, E., Parring, A., Möls. T. (1977) *Tõenäosusteooria ja matemaatiline statistika*, Tallinn Valgus.

[43] Traat, I. (2006) *Matemaatilise statistika põhikursus*, Tartu Ülikooli Kirjastus.

[44] Zar, Jerrhold H. (1984) *Biostatistical Analysis*, Prentice-Hall International Inc.

Aineregister

AIC, 119
alamhulk, 11
ASL, 191
astak, 148
astakkorrelatsioon, 148
asümmeetriaparand, 185, 186

Bayesi valem, 219
beetafunktsioon, 224
beetajaotus, 224, 259
binoomjaotus, 13
bootstrap, 179
bootstrap-meetodid
    BC_a-meetod, 185
    empiiriliselt korduv simuleerimine, 180
    parameetriline *bootstrap*, 181
    t-meetod, 183
Browni liikumine, 252
BUGS, 261

Chapman-Kolmogorovi võrrand, 247
Cooki kaugus, 91

detailse tasakaalu võrrand, 258
determinatsioonikordaja, 86
dispersioonanalüüs
    kahefaktoriline, 97
    ühefaktoriline, 94
dispersiooni test, 49

eeljaotus, 222
eksponentjaotus, 13, 25
eksponentsiaalsete jaotuste pere, 103
ennustaja, 216
erindid, 90
ettepaneku tuum, 258

faktorlaadungite maatriks, 123
FDR, 230
Fisheri
    F-test, 63
    informatsioonikriteerium, 115
    jaotus, 63
Fréchet' jaotus, 146

gammafunktsioon, 221, 225
gammajaotus, 224, 226
geenimarker, 131
Gibbsi
    generaator, 253
    jaotus, 252

valik, 252, 261
Gumbeli jaotus, 146

hii-ruut jaotus, 47
hinnangu
efektiivsus, 15
mõjusus, 15
nihe, 16
nihketus, 15
hälbimus, 113
hüperparameetrid, 244

jacknife-meetod, 176
jaotuse sobimise test, 50
Jenseni võrratus, 178
juhuslik ekslemine, 251
juhuslik protsess, 246
juhuviga, 17
järkstatistikud, 142
järeljaotus, 223

Kendalli τ, 153
keskmine ruutviga, 17
kommunaliteet, 124
korrelatsioonimaatriks, 81
kriitiline väärtus, 37
kvantiil, 27
kvantiili funktsioon, 27

Laplace'i veafunktsioon, 30
logistilised mudelid, 108, 114

Markovi ahela
üleminekumaatriks, 247
üleminekutuum, 250
Markovi ahelad, 246
Metropolis-Hastingsi algoritm, 257
mitmene korrelatsioonikordaja, 81
momendid, 19
mudeli
kirjeldusvõime, 85, 88
maatriks, 78, 94
olulisus, 82, 87
parameetrite olulisus, 84, 87
mütsimaatriks, 91

Neyman-Pearsoni
kriteerium, 236
lemma, 234
nihkeparand, 185
normaaljaotus, 12, 24
normaalne aproksimatsioon, 28
nullhüpotees, 33
nullitriiv, 244

objekt-tunnus maatriks, 75
olulisuse nivoo, 36
olulisustõenäosus, 34
omapära, 124
omavektorid, 130
omaväärtused, 130
otsustuskriteerium, 39, 71

p-value, 35
parameetrite vektor, 12
Pearsoni
χ^2-ruut kriteerium, 51
korrelatsioonikordaja, 80
plaanimaatriks, 76
Poissoni jaotus, 13, 25
PPV, 229

q-value, 232

R2OpenBUGS, 267
Rayleigh' jaotus, 25
regressioonikordajad, 202, 205

sagedustabel, 52, 201

seosefunktsioon, 107, 119
sisukas hüpotees, 33
skalaarkorrutis, 81
skoori test, 118
Speramanni ρ, 151
Stan, 268
standardviga, 17
statistik, 14
statistiline sõltuvus, 208
statistilise mudeli tõestus, 223
statistilise testi võimsus, 39
statsionaarne jaotus, 248
Studenti
    t-jaotus, 58
    t-test, 58
sõltumatus, 202, 213
sõltumatuse test, 52
süstemaatiline viga, 17
šansside suhe, 111, 121

teststatistik, 35
tinglik
    keskväärtus, 202, 213
    ekstreemum, 129
    jaotustihedus, 213
    tõenäosus, 203
tinglikustamine, 221
tsentraalne piirteoreem, 28
Tšuprovi seosekordaja, 207
tõepära
    funktsioon, 22, 234, 241
    suhe, 233, 237
täiendkvantiil, 27
täistinglikud jaotused, 214, 253
täistõenäosuse valem, 219

usaldusnivoo, 27

valim, 10
valimi maht, 11
variatsiooni rida, 142
võimsusfunktsioon, 38
vähimruutude meetod, 78

Waldi test, 117
Weibulli jaotus, 146
Wilcoxoni test, 155

ühistihedus, 213
ühistõenäosus, 201
ühtlane jaotus, 21
üksiktihedus, 213
üksiktõenäosus, 201
üldkogum, 10

Käesoleva õpiku vajadus on tingitud nõudluse kasvust matemaatilise statistika järele Eesti ülikoolides. Enamikel erialadel nii Tallinna Tehnikaülikoolis, Tartu Ülikoolis kui ka Eesti Maaülikoolis on tõenäosusteooria ja matemaatilise statistika põhikursus kohustuslik. Seni aga pole sobivat eestikeelset õpperaamatut selle kursuse matemaatilise statistika osa tarbeks. Käesoleva õpiku üks eesmärke on täita see lünk. Teine eesmärk on olla teatud määral teejuht neile, kellel on vaja rakendada matemaatilist statistikat oma uurimistöödes. Seda eelkõige loodus- ja inseneriteadustes. Õpikus on toodud palju näiteid lahendamaks matemaatilise statistika probleeme tarkvarde MS Excel ja R abil.

Autorist

Margus Pihlak kaitses Tartu Ülikoolis PhD kraadi matemaatilise statistika alal 2007. aastal. Töötanud Tallinna Tehnikaülikooli matemaatikainstituudis ja Eesti Keskkonnaagentuuris. Alates 2008. aastast on õpiku autor Tallinna Tehnikaülikooli matemaatikainstuudis (alates 2016. aastast küberneetikainstituudis) dotsent. Lugenud peamiselt tõenäosusteooria ja matemaatilise statistika põhikursusi, stohhastilise modelleerimise kursust ja inseneristatistikat. Avaldanud teadusartikleid mitmemõõtmelisest statistilisest anlüüsist ja mitteparameetrilise statistika rakendustest keskkonnateadustes.

.

$$f(y) = \exp\left(\ln\left\{ \right.\right.$$

9 789949 832996 >