Peatükk 1

Loeng 1: Pseudojuhuslikud arvud. Lineaarsed kongruentsed generaatorid

Suur osa statistika rakendusi on kas otseselt või kaudselt seotud sõltumatute katsete sooritamisega ja saadud katsetulemuste põhjal mitmesuguste hinnangute leidmisega.

Definitsioon 1 *Arve $x_1, \ldots, x_n$ nimetatakse juhuslikule suurusele X vastavateks juhuslikeks arvudeks, kui nad on saadud juhusliku suuruse X väärtustena juhusliku katse n-kordsel sõltumatul kordamisel.*

Sõltumatuid juhuslikke katseid on raske (ja ilma täiendavate lisaseadmete, nn riistvaraliste juhuarvude generaatoriteta isegi võimatu) arvuti abil piisavalt kiiresti sooritada, seetõttu kasutatakse arvutisimulatsioonides mittejuhuslikke arve, mis imiteerivad sõltumatute katsete tulemusi piisavalt hästi. Katsete sõltumatus on aga väga tugev nõue, mistõttu ei ole juhuslikele arvudele sarnaste mittejuhuslike jadade tekitamine sugugi lihtne ülesanne.

Vaatlema näitena juhtu, kus tahame arvuti abil tekitada ausa täringu viskamisel saadavate tulemustega sarnaselt käituvaid arve. Millised on siis need põhilised omadused, millele peaks meie tekitatud arvujada vastama?

1. Ausa täringu puhul peaks iga konkreetse silmade arvu saamise tõenäosus olema $\frac{1}{6}$. Suurte arvude seaduse kohaselt (kas mäletad, mida see väitis?) peaks seega piisavalt pika visete seeria korral olema iga konkreetse tulemuse esinemiste arvu jagatis visete arvuga küllalt lähedane arvule $\frac{1}{6}$. Udust väljendit "küllalt lähedane" tuleb täpsustada, kuid sellega tegeleme edaspidi. Seega peaks ka meie tekitatud jadal selline omadus olema. Sellest omadusest üksinda aga ei piisa, sest on lihtne tuua näiteid selgelt sobimatutest jadadest, millel on see omadus olemas. Näiteks

$$1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, \ldots$$

2. Eelmises näites tuleb jadas 1 järel alati 2, kuid juhuslike täringuvisete korral võib 1 järel tulla suvaline arv silmi. Tegelikult peaks visete paaride korral esinema kõik arvupaarid $11, 12, \ldots, 66$ sama tõenäosusega. Seega peaks ka meie tekitatud jadal olema selline omadus. Aga ka sellest ei piisa, et jada oleks piisavalt sarnane juhuslike arvudega, sest näiteks jada (seaduspära on ilmne paarikaupa vaadates)

$$111213141516212223 \ldots 656611121314 \ldots$$

korral on see (ja ka eelmine) omadus täidetud.

3. On veel palju omadusi, mis peaks olema täidetud selleks, et arvuti tekitatud jada käituks sarnaselt sellele, mida me juhuslikelt täringuvisetelt ootame (kolmikud peaks olema kõik võrdse tõenäosusega, konkreetse numbri tuleku vahele jäävate visete arvud peavad käituma vastavalt teatavale seaduspärasusele jne).

Eelnevast tulenevalt on simulatsioonideks kasutatavatele juhuslikke arve imiteerivatele arvujadadele mitmesuguseid nõudeid, mille mittetäidetuse puhul võivad tulemused ka suhteliselt lihtsate ülesannete korral olla valed.

Definitsioon 2 *Arve $x_1, \ldots, x_n$ nimetatakse juhuslikule suurusele X vastavateks* pseudojuhuslikeks arvudeks*, kui nad on saadud mingi algoritmi või eeskirja kohaselt ja neid ei õnnestu eristada juhuslikule suurusele X vastavatest juhuslikest arvudest teatud komplekti statistiliste testide abil.*

Eelnev definitsioon ei ole matemaatiliselt päris korrektne, sest kasutatav komplekt statistilisi teste ei ole täpselt määratletud, kuid annab siiski aimu sellest, mida arvutisimulatsioonides pseudojuhuslikke arve kasutades peab silmas pidama. Üldjuhul peatakse pseudojuhuslike arvute tekitamiseks kasutatavat algoritmi (generaatorit) seda paremaks, mida suurema arvu statistiliste testide korral on need eristamatud vaadeldavale juhuslikule suurusele vastavatest juhuslikest arvudest. Konkreetse ülesande lahendamisel arvutisimulatsioonide teel aga oleks hea mõelda, milliseid juhuslike arvude omadusi me lahendamisel kasutame ja kas võib kindel olla, et arvuti poolt genereeritud pseudojuhuslikel arvudel on need omadused olemas.

Kuigi tunnustatud statistikatarkvara poolt kasutatavad pseudojuhuslike arvude generaatorid sobivad enamiku praktiliste ülesannete lahendamiseks, tuleb siiski silmas pidada, et kõigi küsimuste puhul ei pruugi me arvutisimulatsioonidega korrektseid vastuseid saada. Näiteks saab matemaatiliselt tõestada, et lõigus $[0; 1]$ ühtlase jaotusega juhusliku suuruse X korral on sündmuse "X väärtus on ratsionaalarv" toimumise tõenäosuseks 0, kuid arvutisimulatsioonide abil saame tõenäosuseks 1 (kas oskad põhjendada, miks see nii on?). Samuti on näiteks 10000-bitiste täiesti juhuslike jadade genereerimine nii, et kõikvõimalikud jadad esineks sama sageli, üldlevinud generaatorite puhul võimatu (sest erinevate võimalike väljundite arv on väiksem kui kõikide selliste jadade arv).

Järgnevalt vaatleme mõningaid ideid, kuidas pseudojuhuslikke arve genereerida. Vaatleme võimalusi tekitada pseudojuhuslikke reaalarve lõigust [0, 1], mis käituvad nagu juhuslikud arvud jaotusest $U[0, 1]$, sest nendest on sobivate teisendustega võimalik saada mistahes jaotusega (pseudo)juhuslikke arve (tõestame hiljem).

1.0.1 Lineaarne kongruentne generaator (LKG)

Järgnevas tähistame täisarvude x ja $y > 0$ korral x jagamisel y-ga tekkivat jääki kujul $x \bmod y$.

Definitsioon 3 *Etteantud täisarvude m, $0 < a < m$, $0 \leq b < m$ ja $0 \leq x_0 < m$ nimetame eeskirja*

$$x_i := (ax_{i-1} + b) \mod m, \quad u_i := \frac{x_i}{m}$$

abil lõiku $[0, 1]$ kuuluvaid arve u_i, $i \geq 1$ tekitavat algoritmi lineaarseks kongruentseks generaatoriks (tähis LKG(a,b,m)). Kui eelnevas algoritmis kehtib $b = 0$, siis on tegemist multiplikatiivse generaatoriga (inglise keeles ka Lehmer random number generator).

LKG arvutuseeskiri on väga lihtne ja seega võib tunduda uskumatuna, et selliselt midagi juhuslike arvude sarnast tekitada saab. Kõik a, b ja m valikud ei olegi sobilikud: näiteks $a = b = 1$ puhul saame suvalise x_0 ja m korral x_i-de väärtusteks jada $x_0+1, x_0+2, \ldots, m-1, 0, 1, 2, \ldots, m-1, 0, \ldots$, mis kindlasti ei sobi juhuslike arvude rollis kasutamiseks. Lihtne on näha, et LKG väljastab maksimaalselt m erinevat arvu ja seejärel hakkab ennast kordama (alustab uut tsüklit), seega peab m olema piisavalt suur, et arvutisimulatsioonis sellist kordumist ei toimuks. Samas isegi suure m väärtuse korral võib ebaõnnestunud parameetrite valiku korral generaatori tsükli pikkus olla liiga lühike. Näiteks $a = 2^{31},\ m = 2^{32},\ b = 1$ korral saame igast paarisarvulisest x_0 väärtusest lähtuvalt jada $1, 2^{31}+1, 2^{3}1+1, \ldots$ ja igast paaritust x_0 väärtusest lähtuvalt jada $2^{31}+1, 2^{3}1+1, \ldots$. mis samuti ei sobi pseodojuhuslike arvude tekitamiseks. Õnneks võimaldavad arvuteooria tulemused kirjeldada generaatoreid, mis annavad maksimaalse võimaliku arvu erinevaid väärtuseid ning seejärel saab nende sobivust statistiliste testide abil analüüsida.

Teoreem 4 *(Hull-Dobell teoreem, vt [1]) Lineaarse kongruentse meetodi korral saavutatakse tsükli maksimaalne pikkus m parajasti siis, kui on täidetud järgmised tingimused:*

1. *arvude m ja b suurim ühistegur* $\mathrm{SYT}(b, m)$ *on 1;*
2. *kui m jagub p-ga, siis $(a-1)$ jagub p-ga;*
3. *kui m jagub 4-ga, siis $(a-1)$ jagub 4-ga.*

Harjutus 1 *Tõesta eelneva teoreemi esimese tingimuse tarvilikkus maksimaalse perioodi olemasoluks.*

Multiplikatiivse generaatori korral on selge, et kui $x_i = 0$ mingi i korral, siis kõik edasised väärtused on samuti võrdsed nulliga. Seega maksimaalne tsükkel ei saa sisaldada arvu 0 ja selle pikkus ei saa olla seega suurem kui $m-1$.

Teoreem 5 *Multiplikatiivne generaatori tsükkel on maksimaalse pikkusega $m-1$ parajasti siis, kui*

1. *m on algarv ja*
2. *a on algjuur, st iga $m-1$ algteguri p korral $a^{(m-1)/p}$ ei jagu arvuga m*

LKG on oma lihtsuse tõttu on olnud laialdaselt kasutusel, aga praegu ei soovitata neid enam kasutada nt suuremahuliste Monte-Carlo simulatsioonide puhul ning krüpteerimise eesmärgil. Alternatiivid: Mersenne Twister (R vaikegeneraator, vt Wikipedia algoritmi osas), Blum-Blum-Shub generaator (näide võimalikust krüptograafiliselt turvaliseks loetavast generaatorist, vt Wikipedia).

Mõned näited LKG parameetritest, mida on laialdaselt kasutatud (vt Wikipedia, Linear Congruential Generator):

- RANDU generaator (IBM poot kasutusele võetud, tegelikult väga halb generaator): $a = 65539,\ b = 0,\ m = 2^{31}$;
- RANUNI generaator (SAS tarkvaras ole veel üsna hiljuti vaikegeneraator): $a = 397204094,\ b = 0,\ m = 2^{31}$;
- glibc (kasutusel GCC kompilaatoris) $a = 1103515245,\ b = 12345,\ m = 2^{31}$.

Edasises vaatleme mõningaid täiendavaid ideid, kuidas pseodojuhuslikke arve võib genereerida ning seejärel vaatleme, kuidas arvujada vastavust mingist kindlast jaotusest pärinevatele juhuslikele arvudele testida. Neid teste võib kasutada nii empiiriliste andmete käitumise kohta käivate hüpoteeside kontrollimiseks kui ka enda kirjutatud generaatorite korrektsuse kindlakstegemiseks.

Peatükk 2

Loeng 2: Mõningad generaatorite tüübid. Jaotusele vastavuse testimine ühtlase jaotuse korral

Kõigepealt vaatleme mõningaid täiendavaid ideid, kuidas saab pseudojuhuslikke arve genereerida.

2.1 Näiteid generaatoritest

Pseudojuhuslike arvude generaatoreid läheb vaja mitmesuguste seadmete jaoks ja lisaks juhuslikele mudelitele vastavate simulatsioonide läbiviimisele ka näiteks krüptograafias. Erinevad rakendused seavad erinevaid piiranguid, milliseid generaatoreid saab kasutada. Simulatsioonide läbiviimisel on oluline töökiirus, krüptograafiliste rakenduste puhul aga pigem see, et teadaolevate väärtuste komplekti abil ei oleks võimalik kindlaks teha, millised väärtused järgmisena tulevad. Eelnevalt vaadatud LKG-d on head just oma lihtsuse (seega ka töökiiruse) poolest, mistõttu võib neid edukalt kasutada seadmetes, mille protsessori võimsus on üsna väike. Samas isegi küllalt vähevõimsate protsessorite töökiirus on kaasajal nii kõrge, et protsessori bittide arvule vastava mooduli m väärtusega LKG-de puhul võib tekkida probleem, et ühe simulatsiooni käigus kasutab generaator kõik väärtused ära ja alustab otsast peale. See aga kindlasti ei ole kooskõlas sellega, kuidas juhuarvud peaksid käituma. Osutub, et sellest probleemist on võimalik üle saada lihtsate generaatorite kombineerimise teel. Toome ühe näite.

2.1.1 Whichman-Hill generaator

Järgnev kirjeldus põhineb Wikipedia artiklil `https://en.wikipedia.org/wiki/Wichmann%E2%80%93Hill`

Pseodujuhuslikud arvud u_i genereeritakse järgnevalt:

$$\begin{aligned} x_i &= (171 \cdot x_{i-1}) \mod 30269, \\ y_i &= (172 \cdot y_{i-1}) \mod 30307 \\ z_i &= (170 \cdot z_{i-1}) \mod 30323, \\ u_i &= \left(\frac{x_i}{30269} + \frac{y_i}{30307} + \frac{z_i}{30323}\right) \mod 1 \end{aligned}$$

Selle generaatori perioodiks on 6953607871644 ja arvutusi saab edukalt läbi viia isegi ainult

16-bitisel protsessoril. Samas tasub teada, et midagi sisuliselt erinevat selline kombineerimine ei anna, sest see generaator on samaväärne LKG-ga juhul $a = 16555425264690$, $b = 0$, $m = 27817185604309$. Viimast otse realiseerida nii, et töökiirus ei kannataks, on aga suhteliselt tülikas.

2.1.2 Mersenne Twister

Lineaarsete kongruentsete generaatorite puhul saab näidata, et kui moodustatakse punkte järjestikustest väärtustest 2, 3 jne dimensioonilises ruumis, siis need paiknevad kõik paralleelsetel hüpertasanditel, kusjuures vajaminevate hüpertasandite arv langeb üsna kiiresti dimensiooni kasvades [3]. See tähendab, et kui huvipakkuva juhusliku suuruse genereerimiseks läheb vaja palju sõltumatuid juhuslikke arve, siis võib LKG-del baseeruvate generaatorite korral tekkida arvutatavates keskmistes küllalt suur viga. Et selliseid probleeme vältida, pakkusid Makoto Matsumoto ja Takuji Nishimura 1997a välja algoritmi nimega Mersenne Twister, mis on praegusel hetkel vaikegeneraatoriks paljudes tarkvarasüsteemides, sh R, Python, SAS jne. Generaatori detailne algoritm on leitav Wikipediast (`https://en.wikipedia.org/wiki/Mersenne_Twister`), aga mingi aimuse võib saamiseks toome järgmise juhuarvu arvutamise tehte ilma kõikide detailideta:

$$x_i = x_{i-m} \oplus (g(x_{i-n}, x_{i-n+1})A),$$

kus g paneb kokku uue w-bitise arvu kahe varem genereeritud sama pika arvu erinevatest osadest ning (tulemust vaatleme reavektorina) ning A on spetsiaalse struktuuriga maatriks (ehk lineaarteisendus) ning $\oplus$ tähistab bitiviisilist välistava või tehet, mis on arvutuslikus mõttes väga efektiivselt teostatav. Nihked m ja n on mainitud algoritmis valitud nii, et saadava generaatori tsükli pikkus on nn Mersenne algarv $2^{19937} - 1$.

2.1.3 Blum-Blum-Shub generaator

Kuna krüpteerimisel kasutatavatel juhuarvude generaatorite puhul on väga oluline see, et mingi komplekti väljastatud arvude baasil ei oleks realistlik tuletada eeskirja järgmiste väljastatavate juhuarvude tekitamiseks isegi juhul, kui kasutatav algoritm on teada, ei sobi selleks eelnevalt kirjeldatud generaatorid. Matemaatiliselt ei pruugigi sobivate generaatorite kirjeldused väga keerulised olla, aga nende abil väärtuste genereerimine on tavaliselt küllaltki aeganõudev ja seetõttu selliseid generaatoreid ei kasutata tavaliselt arvutisimulatsioonides. Üheks selliseks generaatoriks on Lenore Blum, Manuel Blum ja Michael Shub'i poolt 1986a loodud nn Blum-Blum-Shub generaator, mille arvutuseeskiri on järgmine:

$$x_i = x_{i-1}^2 \mod m,$$

kus $m = p \cdot q$, p ja q on suured raskestiaimatavad algarvud, mille jääk jagamisel 4-ga on 3 (nn Blum'i algarvud). Näiteks sobivad: p = 12676506002282294014967039815l9, q = 126765060022822940149670431835 9 (kuid muidugi iga rakenduse korral tuleb need juhuarvud uuesti genereerida). Generaatori väljundiks on bitijada $u_i = x_i \mod 2$, st bitijada, millest saab tekitada soovitud bittide arvuga pseudojuhuslikke arve. Nagu näha, tuleb iga biti genereerimiseks teha tehteid väga suurte arvudega ja seetõttu on see generaator üsna aeglane.

2.2 Generaatorite väljundite sobivuse kontrollimine

Tahame kontrollida, et generaatori väärtused käituvad nagu soovitud jaotusele vastavad juhuslikud arvud. Selleks kirjeldame juhuarvude käitumist valimis mitmesuguste statistikute (valimifunktsioonide) abil ja uurime, kas generaatori väljund (või ka lihtsalt vaatluste jada väljund) annab oodatavaid (st juhuarvude korral piisavalt suure tõenäosusega tekkivaid) tulemusi. Vaatleme kahte tüüpi teste - jaotuse kontrollimise testid (kas väärtused paiknevad reaalteljel nii, nagu soovitava jaotusega juhusliku suuruse väärtused peaksid käituma) ja sõltumatuse testid (kas väärtused järgnevad üksteisele nii nagu sõltumatute katsete tulemused peaks käituma).

2.3 Testimine ühe sündmuse abil

Varasematest kursustest teame, et fikseeritud sündmuse A toimumiste arv n sõltumatus katses on binoomjaotusega $B(n,p)$, kus $p = P(A)$ on A toimumise tõenäosus ühel katsel. Kui meil on teada näiteks juhusliku suuruse X jaotusfunktsioon F_X ja A on kujul

$$A = \{a < X \leq b\},$$

siis jaotusfunktsiooni definitsioonist järeldub võrdus

$$P(A) = F_X(b) - F_X(a).$$

Kui me sooritame sõltumatuid katseid n korda ja loeme kokku A toimumiste arvu N_A, siis eelneva põhjal $N_A \sim B(n,p)$. Enamasti on põhjust etteantud jaotusele vastavuses kahelda siis, kui sündmuse A toimumiste arv erineb oluliselt oodatavast keskväärtusest $E(N_A) = np$. Kuna generaatori ekslikult mittesobivaks lugemine toob kaasa lisatööd (tuleb otsida vigu, leida parem generaator vms), ei taha me enamasti liiga kergekäeliselt katsetulemusi kahtlaseks lugeda. Seega valime küllalt väikese α korral sellised arvud x_1 ja x_2, et vaadeldavale jaotusele vastavate juhuslike arvude korral kehtib

$$P(-x_1 < N_A - np \leq x_2) \geq 1 - \alpha$$

ning loeme kahtlaseks tulemused, mis jäävad väljaspoole neid piire. Sobiv valik x_1 ja x_2 jaoks vastab võimalikut väikestele positiivsetele arvudele, mille korral kehtivad

$$P(N_A \leq np - x_1) \leq \frac{\alpha}{2}$$

ja

$$P(N_A \leq np + x_2) \leq 1 - \frac{\alpha}{2},$$

mis puhul eelnev tingimus on samaväärne katsetulemuste põhjal leitud N_A väärtuse jäämisega vaadeldava binoomjaotuse $\frac{\alpha}{2}$ ja $(1 - \frac{\alpha}{2})$-kvantiilide vahele.

Täiendavaks testimise võimaluseks on eelnevat tegevust (n väärtuse korral vaadeldava sündmuse toimumiste arvu leidmist ja selle teoreetilisele jaotusele vastavuse kontrollimist) korrata mitu korda. Seda võime käsitleda ühe keerulise katse teostamisena (kus tulemuseks on 1, kui toimumiste arv jäi valitud kvantiilide vahele ja 0 siis, kui ei jäänud). On selge, et vaadeldavale jaotusele vastavate juhuarvude korral peaks aeg-ajalt (tõenäosusega $1-\alpha$) jääma katse põhjal saadud väärtus eelnevalt defineeritud piiridest välja. Täpsemalt, vaadeldavale jaotusele vastavate juhuslike arvude korral peaks sellise katse puhul saadavate ühtede arv olema jällegi binoomjaotusega $B(m,\alpha)$, kus m on testi kordamiste arv. Sarnaselt eelnevale saame kontrollida, kas tulemused on kooskõlas teoreetilise jaotusega.

Viimast ideed konstrueerida täendav test mingi konkreetse testi korduva teostamise abil saame rakendada ka kõigil järgnevatel juhtudel.

2.4 Diskreetsele jaotusele vastavuse testimine χ^2-testiga

Kui vaadelda katset, millel on m võimalikku erinevat ja vastastikku välistavat katsetulemust, siis on see katse tõenäosuslikus mõttes täielikult kirjeldatud võimalike katsetulemuste ja nende tõenäosuste paaridega. Sel juhul on loomulik idee kontrollida, kas katseseeriale vastavas sagedustabelis on iga katsetulemuse esinemiste arv kooskõlas tema tõenäosusega, kuid seda kooskõlas olemise tingimust tuleb täpsustada. Kuna iga sellise katse puhul saame võimalikud katsetulemused ära nummerdada ning katsetulemusi samastada juhusliku suurusega, mille väärtuseks on i kui katsetulemuseks on i-s element nummerdatud katsetulemuste hulgast, siis võime öelda, et järnevas vaatleme diskreetsete juhuslike suuruste etteantud jaotusele vastavuse testimist.

Loomulik on jaotusega kooskõla uurimisel vaadelda katsetulemuste esinemiste arvu erinevusi oodatavatest esinemiste arvudest (vastavate juhuslik suuruste keskväärtustest) np_i, kusjuures väga suured erinevused on mõistlik lugeda kahtlasteks. Samas on selge, et erinevusi keskmistest ei ole hea vaadata ükshaaval, sest vähegi suurema sündmuste komplekti korral on ka õigest jaotusest pärinevate katsetulemuste korral küllalt suur tõenäosus, et mõne sündmuse esinemissagedus erineb üsna palju oodatavast. Seega on hea vaadelda mingit summaarset oodatavatest sagedustest erinevuse mõõdikut, mille jaotus (või vähemalt asümptootiline jaotus piisavalt suure katsete arvu korral) on vaadeldavast jaotusest pärinevate juhuslike arvude korral teada. Üks selline mõõdik on χ^2-statistik.

2.4.1 Hii-ruut test

Vaatleme juhuslikku katset, millel on i erinevat vastastikku välistavat katsetulemust tõenäosustega $p_i,\ i = 1, \ldots, m$. Olgu katseseeria põhjal (ehk n genereeritud väärtuse põhjal) leitud toimumiste arvud $n_i,\ i = 1, \ldots, m$. **Pearsoni χ^2-statistiku** väärtus on sel juhul defineeritud valemiga

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

On teada, et piiril $n \to \infty$ on see statistik χ^2 jaotusega vabadusastmete arvuga $m - 1$. *NB! Enamasti (vt näiteks Wikipedia testile pühendatud artikkel inglisekeelses versioonis) loetakse, et piirjaotuse kasutamine on õigustatud ainult siis, kui n on nii suur, et $np_i \geq 5\ \forall i$.*

Testi rakendamisel on nullhüpoteesiks katsetulemuste vastamine vaadeldavale jaotusele ning katsetulemused loetakse vaadeldavale jaotusele mittevastavaks, kui statistiku väärtus on liiga suur, st kuulub piirkonda $[q_{1-\alpha}, \infty)$, kus $q_{1-\alpha}$ on vastava hii-ruut jaotuse $(1-\alpha)$-kvantiil ja α on valitud olulisuse nivoo. Samaväärne otsustuseeskiri statistiku p-väärtuse kaudu on järgmine: kui see on väiksem kui meie poolt valitud olulisuse nivoo α, siis loeme nullhüpoteesi ümberlükatuks; vastasel korral aga jääme nullhüpoteesi juurde (st generaatori testimise kontekstis vaadeldavad katsetulemused ei anna piisavalt alust generaatori õigsuses kahtlemiseks).

Hii-ruut testi baasil saab konstrueerida teste igasugu jaotuste testimiseks, teisendades katsetulemused sobival moel lõplikuks arvuks erinevateks väärtusteks. Mõningatel juhtudel on saadud tuletatud testil oma nimi. Sellist teisendust saab teostada näiteks fikseerides on m vastastikku välistavat sündmust $A_1, \ldots, A_m$, mille tõenäosused vaadeldava jaotuse korral on $p_1, \ldots, p_m$ ($p_i > 0\,\forall i$) ja mis katavad ära kõik võimalused (täissüsteem). Näiteks juhusliku suuruse X korral võime vaadelda sündmuseid $A_i = \{x_{i-1} < X \leq c_i\}$, kus

$$-\infty = c_0 < c_1 < \ldots < c_{m-1} < c_m = \infty$$

ja punktid $c_i,\ i = 1, 2, \ldots, m-1$ on valitud nii, et kõik sündmused on vaadeldava jaotuse puhul võimalikud (st positiivse tõenäosusega). Ühtlasele jaotusele vastavuse testimise puhul on sellel testile oma nimi - **sageduste test**, kui sündmuseid defineerivateks reaaltelje jaotuspunktideks on $c_i = \frac{i}{m},\ i = 1, 2, \ldots, m$.

Peatükk 3

Loeng 3: Pideva jaotuse väärtuste paiknemise testimine. Sõltumatuse testimine

3.1 Empiirilise ja teoreetilise jaotusfunktsiooni võrdlemisel põhinevad testid

Pideva või lõpmatu arvu erinevate väärtustega diskreetse juhusliku suuruse testimisel eelneva ideega on raske otsustada, milliseid piirkondi valida. Et selle küsimusega mitte oma pead vaevata, on loomulik vaadelda korraga komplekti lõpmatult paljudest piirkondadest kujul $A_x = \{X \leq x\}$, kus x reaalarv. Sellisesse piirkonda sattumise tõenäosus vaadeldava jaotuse korral on $F_X(x)$ (kus F_X on X jaotusfunktsioon). Valimi $x_1, \ldots, x_n$ korral piirkonda A_x kuuluvate vaatluste suhteline sagedus vastab **empiirilisele jaotusfunktsioonile**

$$F_n(x) := \frac{\#\{i \mid x_i \leq x\}}{n},$$

kus $\#B$ tähistab hulga B elementide arvu. Seega konkreetse x korral iseloomustab vahe $F_n(x) - F(x)$ sündmuse A_x empiirilise esinemissageduse erinevust teoreetilisest tõenäosusest ning suur erinevus võiks anda alust kahtlusele, et andmed ei vasta vaadeldava jaotusega juhuslikele arvudele. Sündmuste korraga vaatlemiseks aga on vaja kasutada mingit mõõdikut, mis kõikidele reaalarvudele x vastavad erinevused kokku võtaks ning selleks on mitmeid võimalusi, millest vaatame järgnevalt kahte.

Kõigepealt tõestame ühe tulemuse, mis on järgnevate testide tööpõhimõttest arusaamisel äärmiselt oluline.

Teoreem 6 *Olgu F pidev jaotusfunktsioon ning olgu X sellele vastav juhuslik suurus. Siis $Y = F(X)$ on ühtlase jaotusega $U(0,1)$*

Tõestus. Näitame, et juhusliku suuruse Y jaotusfunktsioon F_Y on ühtlase jaotuse jaotusfunktsioon, st avaldub kujul

$$F_Y(y) = \begin{cases} 0, & \text{kui } y <= 0, \\ y, & \text{kui } 0 < y < 1 \\ 1 & \text{kui } y \geq 1. \end{cases}$$

Selleks paneme tähele, et

$$F_Y(y) \underset{def}{=} P(\{Y \leq y\}) = P(\{F(X) \leq y\}) = P(\{X \in A_y\}),$$

kus hulk A_y on selliste reaalarvude hulk, mille korral F väärtus ei ole suurem y väärtusest, ehk

$$A_y = \{x \in \mathbb{R} : F(x) \leq y\}.$$

Vaatleme eraldi juhte $0 < y < 1$, $y \leq 0$ ja $y \geq 1$

juht $0 < y < 1$: Jaotusfunktsiooni omadusest 3 (Vaata teoreem 24) ja F pidevusest järeldub, et leidub selline reaalarv a_y, mille korral $F(a_y) = y$, seega A_y ei ole tühihulk. Jaotusfunktsiooni monotoonse kasvamise omadusest (Teoreem 24, omadus 2) järeldub võrratusest $x \leq a_y$ võrratus $F(x) \leq y$, seega

$$(-\infty, a_y] \subset A_y.$$

Tõenäosuse monotoonsuse omadusest (Teoreem 21, omadus 3) järeldub seega võrratus

$$y = F(a_y) = P(\{X \in (-\infty, a_y]\}) \leq P(\{X \in A_y\}) = F_Y(y).$$

Veendume ka vastupidise võrratuse kehtimises. Olgu $\varepsilon > 0$ selline reaalarv, mille korral $y+\varepsilon < 1$. Jällegi saame F pidevuse tõttu leida sellise $a_{y+\varepsilon}$, et $F(a_{y+\varepsilon}) = y+\varepsilon$. Kuna iga $x \geq a_{y+\varepsilon}$ korral kehtib F monotoonse kasvamise tõttu $F(x) \geq y + \varepsilon > y$, siis $A_y \subset (-\infty, a_{y+\varepsilon}]$. Seega jällegi tõenäosuse monotoonsuse põhjal

$$F_Y(y) = P(\{X \in A_y\}) \leq P(\{X \in (-\infty, a_{y+\varepsilon}]\}) = F(a_{y+\varepsilon}) = y + \varepsilon.$$

Piiril $\varepsilon \to 0$ saame siit $F_Y(y) \leq y$ ning eelnevat vastupidist võrratust arvestades olema näidanud, et kehtib võrdus

$$F_Y(y) = y, \ 0 < y < 1.$$

juht $y \leq 0$: iga reaalarvu $\varepsilon \in (0, 1)$ korral kehtib jaotusfunktsiooni monotoonsuse tõttu vaadeldaval juhul võrratus $F_Y(y) \leq F_Y(\varepsilon) = \varepsilon$. Piiril $\varepsilon \to 0$ saame siit $F_Y(y) \leq 0$. Jaotusfunktsiooni mittenegatiivsusest järeldub siit, et kehtib võrdus

$$F_Y(y) = 0, \ y \leq 0.$$

juht $y \geq 1$: Kuna iga jaotusfunktsioon rahuldab tingimust $F(x) \leq 1 \ \forall x \in \mathbb{R}$, siis A_y on kogu reaaltelg ja seetõttu X väärtuse sattumine hulka A_y on kindel sündmus. Järelikult $y \geq 1$ korral

$$F_Y(y) = 1.$$

Sellega on teoreemi väide tõestatud.□

3.1.1 Kolmogorov-Smirnovi test

Kolmogorov-Smirnovi test võimaldab otsustada, kas valim pärineb etteantud pidevale jaotusfunktsioonile F vastavast jaotusest. Testi nullhüpüteesiks on see, et valim koosneb jaotusfunktsioonile F vastavatest juhuslikest arvudest (sõltumatute katsete tulemustest). Alternatiivne hüpotees H_1 on see, et valim ei koosne jaotusfunktsioonile F vastavatest juhuslikest arvudest. Otsustamiseks kasutatakse statistiku

$$K_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

valimi põhjal arvutatud väärtuse p-väärtust ehk tõenäosust, et jaotusfunktsioonile F vastavate juhuslike arvude korral saadakse vähemalt sama suur väärtus, kui vaadeldava valimi

korral. Kui p-väärtus on väiksem, kui meie valitud olulisuse nivoo α, loeme tõestatuks alternatiivse hüpoteesi.

Veendume, et statistiku jaotus nullhüpoteesi kehtimisel ei sõltu vaadeldavast jaotusfunktsioonist F. Tähistagu $x_{(i)}$ valimi suuruselt i-ndat elementi, kusjuures defineerime lisaks

$$x_{(0)} = -\infty,\ x_{(n+1)} = \infty.$$

Kuna

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \max_{i=0,\dots,n} \sup_{x\in[x_{(i)},x_{(i+1)})} |F_n(x) - F(x)|,$$

siis leiame järgmiseks supreemumid üle osalõikude.

Empiirilise jaotusfunktsiooni definitsioonist järeldub, et

$$F_n(x) = \frac{i}{n},\ i = 0,\dots,n,\ x \in [x_{(i)},\ x_{(i+1)}),$$

seega

$$\sup_{x\in[x_{(i)},x_{(i+1)})} |F_n(x) - F(x)| = \sup_{x\in[x_{(i)},x_{(i+1)})} |F(x) - \frac{i}{n}|.$$

Funktsiooni F mittekahanemisest järeldub, et

$$F(x_{(i)}) - \frac{i}{n} \le F(x) - \frac{i}{n} \le F(x_{(i+1)}) - \frac{i}{n},\ x \in [x_{(i)}, x_{(i+1)}),$$

kusjuures F pidevuse tõttu

$$\sup_{x\in[x_{(i)},x_{(i+1)})} F(x) - \frac{i}{n} = F(x_{(i+1)}) - \frac{i}{n}.$$

Analüüsides juhte $F(x_{(i+1)}) < \frac{i}{n}$, $F(x_{(i)}) < \frac{i}{n}$, $F(x_{(i+1)} \ge \frac{i}{n})$, $F(x_{(i)}) \ge \frac{i}{n}$, saame kõigil juhtudel, et

$$\sup_{x\in[x_{(i)},x_{(i+1)})} |F_n(x) - F(x)| = \max\{\frac{i}{n} - F(x_i), F(x_{(i+1)}) - \frac{i}{n}\}.$$

Seega kehtib võrdus

$$K_n = \max_{i=1,\dots,n} \{F(x_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F(x_{(i)})\}.$$

Kuna aga $F(X)$ on ühtlase jaotusega $U(0,1)$ juhuslik suurus, siis $y_i = F(x_i)$ on nullhüpoteesi korral sõltumatu valim jaotusest $U(0,1)$, $y_{(i)} = F(x_{(i)})$ on selle valimi suuruselt i-s element ning seega statistiku jaotus ei sõltu vaadeldavast jaotusfunktsioonist, vaid on leitav statistiku käitumise uurimisega jaotusest $U(0,1)$ pärineva valimi korral.

3.1.2 Oomega-ruut test

Erinevust $F_n(x) - F(x)$ saab mõõta ka teisiti. Kui X on pidev juhuslik suurus jaotusfunktsiooniga F ja tihedusfunktsiooniga f, siis Cramer - von Mises test ehk oomega-ruut test põhineb järgmisel F ja F_n erinevust mõõtval statistikul:

$$\omega_n^2 := n \int_{-\infty}^{\infty} (F(x) - F_n(x))^2 f(x) dx.$$

Saab näidata (vt Kollo õpik, lk 35-36), et kehtib

$$\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_{(i)}) - \frac{i-0.5}{n} \right)^2,$$

seega jällegi ei sõltu statistiku jaotus nullhüpoteesi kehtimisel vaadeldavast jaotusfunktsioonist F, vaid on leitav ühtlasest jaotusest pärineva valimi analüüsimise teel. Otsustamise eeskiri (hüpoteesid) on samad, kui KS testi puhul.

3.2 Sõltumatuse testimine

Eelnevad testid arvestavad ainult valimis olevate arvude paiknemist reaalteljel, kuid ei arvesta nende järgnevusega seotud infot. Näiteks saaksime samad testitulemused, kui eelnevalt sorteeriksime valimi kasvamise järjekorras. Samas sõltumatute katsete tulemuste puhul on oluline ka järgnevus, st eelmiste väärtuste teadmine ei tohi anda mingit infot järgmiste väärtuste kohta ja kui generaator väljastab väärtuseid kasvavas järjekorras, siis ei vasta need sõltumatute katsete tulemustele.

Järgnevalt vaatleme teste, mis arvestavad nii soovitud jaotusega seotud tõenäosuseid kui ka sõltumatute katsete puhul kehtivaid järjestikuste katsetulemustega seotud omadusi.

3.2.1 Sõltumatute paaride test

Sõltumatute katsete põhiomaduseks on see, et neile vastavate sündmuste koos toimumise tõenäosus on vastavate sündmuste tõenäosuste korrutis. Sõltumatute paaride testi korral vaadeldakse järjestikuste väärtuste paare ehk liitkatset, kus iga kord tekib arvupaar. Testi sooritamise algoritm on järgmine:

1. Fikseerime m vastastikku välistavat sündmust $A_1, \ldots, A_m$, mille tõenäosused vaadeldava jaotuse korral on $p_1, \ldots, p_m$ ja mis katavad ära kõik võimalused (täissüsteem)
2. Igale katsetulemusele seame vastavusse sündmuse numbri, mis sellel katsel toimus. Saame numbritest $1, .., m$ koosneva jada $k_1, k_2, \ldots, k_n$.
3. Vaatleme paare $(k_1, k_2), (k_3, k_4), \ldots$. Kui tegemist on sõltumatute katsetega, siis paari (i, j) saamise tõenäosus igal liitkatsel on $p_i \cdot p_j$
4. Kontrollime χ^2-testiga, kas arvupaaride $(i, j),\ i, j \in \{1, \ldots, m\}$ esinemissagedused on kooskõlas nende saamise tõenäosusega.

Selgituseks: kui meil on $2n$ väärtusest koosnev valim, siis võime nende põhjal tekitatud arvupaare $(k_{2i-1}, k_{2i}),\ i = 1, \ldots, n$ vaadelda kui valimit n liitkatse tulemustest, mille korral sündmused A_{ij} ="liitkatse puhul saadakse paar (i,j)"moodustavad täissüsteemi. Seega on tegemist just sellise juhuga, mille korral saab hii-ruut testi rakendada.

Kui testi p-väärtus on väiksem kui meie olulisuse nivoo, siis loeme tõestatuks, et katsetulemused ei vasta vaadeldava jaotusega juhuslikele (või pseodojuhuslikele) arvudele. Mittevastavuse põhjuseks võib olla kas etteantud jaotusele mittevastamine (tõenäosused ei klapi) või siis see, et katsetulemused ei ole sõltumatud.

Märkus: χ^2-testil on olemas versioon, mis kontrollib ainult sõltumatust. Kontroll seisneb andmete põhjal sündmuste $A_i,\ i = 1, \ldots, m$ tõenäosuste hindamises ja kontrollis, kas arvupaaride (i, j) esinemissagedused on kooskõlas hinnatud tõenäosuste korrutisega hii-ruut statistiku abil, kus teoreetilised tõenäosused on asendatud hinnatud tõenäosustega.

Kuna siin kursusel me kontrollime teadaolevale jaotusele vastamist, siis seda testi versiooni me ei kasuta.

3.2.2 Vahemike test

Vaatleme järgmist tegevust:

- Fikseerime mingi sündmuse A, mille toimumise tõenäosus ühel katsel on p
- teostame katseid, kuni sündmus A toimub, paneme kirja selleks kuluvate katsete arvu
- teostame jälle katseid, kuni sündmus A toimub, paneme kirja selleks kuluvate katsete arvu
- jne

Vaadeldes valimit kui sellise tegevuse teostamisel kirjapandud katsetulemusi, saame teha kindlaks, mitu eelpool kirjeldatud katseseeriat teostati (viimane võib olla poolik ja seda me ei arvesta) ning samuti ka iga seeria korral saadud tulemuse (vajaläinud katsete arvu). Seega saame esialgsest valimist uue (lühema) valimi liitkatsete tulemustega. Teame, et katsete arv kuni fikseeritud sündmuse toimumiseni on geomeetrilise jaotusega $Geom(p)$, seega uus valim peaks vastama valimile jaotusest $Geom(p)$. Fikseerides nüüd uue valimi korral ℓ sündmust kujul "A tulekuks kulub i katset", kus $i = 1, 2, \ldots, \ell - 1$ ja "A tulekuks kulub vähemalt ℓ katset", saame hii-ruut testiga kontrollida tulemuste vastavust jaotusele $Geom(p)$. (NB! Leia geomeetrilise jaotuse definitsiooni põhjal viimati defineeritud sündmuse tõenäosus!)

3.2.3 Seeriate test

Vahemike test nõuab ühe sündmuse fikseerimist. Samas võime vaadelda suurema arvu sündmuste korral seda, kui pikad on sama sündmuse järjestikustel katsetel kordumiste jadad. Võime mõelda järgmisest tegevusest:

- Fikseerime m vastastikku välistavat võrdvõimalikku sündmust $A_1, \ldots, A_m$
- Teeme ühe katse, paneme kirja selle korral toimunud sündmuse numbri
- Teeme katseid, kuni tuleb kirjapandud numbriga sündmusest erinev sündmus. Fikseerime selleks kulunud katsete arvu (ehk kirjapandud numbriga sündmuste tulemise seeria pikkuse) ja paneme kirja viimasel katsel saadud sündmuse numbri
- Kordame eelmises kirjeldatud tegevusi soovitud arvu kordi.

Kuna iga seeria korral teeme sõltumatuid katseid, kuni toimub kirjapandud numbriga sündmusest erinev sündmus, siis on igal katsel seeria lõppemise tõenäosus $\frac{m-1}{m}$ ja seega on vajaminevate katsete arv jaotusega $Geom(\frac{m-1}{m})$. Kui vaadelda algset valimit kui sellise tegevuse käigus kirjapandud katsetulemusi, saame nende põhjal kindlaks teha teostatud katsevoorude arvu ja nende käigus leitud seeriate pikkused. Kui esialgne valim koosneb vaadeldavale jaotusele vastavatest juhuslikest arvudest, siis leitud seeriate pikkused moodustavad valimi jaotusest $Geom(\frac{m-1}{m})$. Saadud seeriate pikkuste valimi vastavust geomeetrilisele jaotuse saame testida samamoodi nagu vahemike testi puhul

3.2.4 Pokker test

Pokker-testi algoritm on järgmine:

1. Fikseerime m vastastikku välistavat võrdvõimalikku sündmust $A_1, \ldots, A_m$.
2. Igale katsetulemusele seame vastavusse sündmuse numbri, mis sellel katsel toimus.
3. Vaatleme toimunud sündmuste numbritele vastavas jadas viisikuid $(k_1, k_2, k_3, k_4, k_5), \ldots$.
4. Loendame erinevate mustrite esinemisi, mustriteks on: kõik erinevad, üks paar, kaks paari, kolmik, paar ja kolmik, nelik, viisik.
5. Kasutame χ^2-testi mustrite esinemissageduste võrdlemisel teoreetiliste sagedustega.

Mustrite teoreetilised esinemissagedused saab leida kombinatoorika valemeid kasutades (vt Kollo õpikust [2]).

Lisaks eelmainitud testidele on välja töötatud suur hulk täiendavaid teste, mille abil uusi generaatoreid kontrollitakse. Näitena võib tuua George Marsaglia poolt väljatöötatud testikomplekti "Diehard Test Suite" ja TestU01 testikomplekti, mille kohta saab infot veebilehelt `http://simul.iro.umontreal.ca/testu01/tu01.html`

Peatükk 4

Loeng 4: Jaotusfunktsiooni pööramise meetod juhuarvude genereerimiseks

4.1 Etteantud jaotusele vastavate pseudojuhuslike arvude genereerimine

Järgnevalt eeldame, et me oskame genereerida ühtlasele jaotusele $U(0,1)$ vastavaid pseudojuhuslikke arve või meil on füüsiline generaator, mis väljastab sellele jaotusele vastavaid juhuslikke arve. Vaatleme võimalusi, kuidas nende abil saab tekitada suvalisele jaotusele vastavaid (pseudo)juhuslikke arve.

4.1.1 Jaotusfunktsiooni pööramise meetod

Kuna eelneva põhjal teame, et suvalise pideva jaotusfunktsiooniga juhuslikust suurusest saame sellele jaotusfunktsiooni rakendades ühtlase jaotusega $U0,1)$ juhusliku suuruse, siis võib loota, et jaotusfunktsiooni pöördfunktsiooni abil saab ühtlasest jaotusest soovitud jaotusega juhusliku suuruse.

Vaatleme esialgu lihtsat erijuhtu, kus vaadeldava jaotuse jaotusfunktsioon F on rangelt kasvav kogu reaalteljel. Sel juhul on sel funktsioonil ka vahemikul (0, 1) defineeritud pöördfunktsion F^{-1}. Olgu $U \sim U(0,1)$ ning defineerime juhusliku suuruse X kujul $X = F^{-1}(U)$. Siis funktsiooni F range kasvamise tõttu on iga $x \in \mathbf{R}$ korral võrratus $X \leq x$ samaväärne võrratusega $U = F(X) \leq F(x)$ ning kuna U on vahemikus (0, 1) ühtlase jaotusega, siis $P(U \leq F(x)) = F(x)$, st $F_X(x) = P(X \leq x) = F(x)$ ehk X jaotusfunktsiooniks on F.

Osutub, et selline moodus ühtlase jaotusega juhuslikust suurusest etteantud jaotusfunktsiooniga F juhusliku suuruse saamiseks töötab alati, kui pöördfunktsiooni mõistet sobivalt üldistada juhtudele, kus tavalist pöördfunktsiooni ei leidu.

Definitsioon 7 *Olgu $F : R \to [0,1]$ mingi tõenäosusjaotuse jaotusfunktsioon. Siis selle üldistatud pöördfunktsiooniks nimetatakse funktsiooni $F^- : (0,\ 1) \to \boldsymbol{R}$, mis on defineeritud võrdusega*

$$F^-(y) = \inf\{a\ :\ F(a) \geq y\},\ y \in (0\ ,\ 1).$$

Jaotusfunktsiooni üldistest omadustest järeldub, et nii defineeritud pöördfunktsioon alati korrektselt defineeritud (väärtused on üheselt määratud reaalarvud). (Veendu selles!)

Eelneva definitsiooni üle järele mõeldes on küllalt lihtne veenduda, et juhul, kui vaadeldava y korral võrrandil $F(x) = y$ on ühene lahend $x = x_y$, siis $F^- y = x_y$, seega pöördfunktsiooni olemasolul langevad üldistatud pöördfunktsioon ja tavaline pöördfunktsioon kokku. Kui selliseid lahendeid on aga mitu, siis pöördfunktsiooni väärtuseks võetakse lahendite hulga alumine raja ning kui lahendeid pole, siis on pöördfunktsiooni väärtuseks see x väärtus, mille korral funktsiooni graafik "hüppab"läbi taseme y.

Järgnev tulemus lihtsustab meil põhitulemuse tõestamist.

Lemma 8 *Olgu $F : R \to [0,1]$ mingi jaotuse jaotusfunktsioon. Siis selle üldistatud pöörfunktsiooni korral on võrratus $F^-(y) \leq x$ hulgal $\{(x,y) : y \in (0,1),\ -\infty < x < \infty\}$ samaväärne võrratusega $y \leq F(x)$.*

Tõestus. Samaväärsuse näitamiseks on vaja näidata, et esimesest võrratusest järeldub teine ning et teisest järeldub esimene.

a) Olgu x ja $y \in (0,1)$ sellised, et kehtib võrratus $F^-(y) \leq x$. Funktsiooni F^- definitsiooni kohaselt leidub siis selline monotoonselt kahanev jada a_n, et $\lim_{n\to\infty} a_n \leq x$ ning $F(a_n) \geq y$. Defineerides $b_n = \max(a_n, x)$ saame uue monotoonselt kahaneva jada, mis funktsiooni F mittekahanevuse tõttu rahuldab samuti tingimust $F(b_n) \geq y\ \forall n$ ning samuti kehtib $\lim_{n\to\infty} b_n = x$. Kuna F on paremalt pidev, kehtib seega ka
$$F(x) = \lim_{n\to\infty} F(b_n) \geq y.$$

b) Olgu x ja y sellised, et kehtib võrratus $y \leq F(x)$. Siis
$$x \in \{a : F(a) \geq y\}$$
ning seetõttu
$$F^-(y) = \inf\{a\ :\ F(a) \geq y\} \leq x.$$

Lemma on tõestatud. □

Üldistatud pöördfunktsiooni abil saab esitada järgmise olulise tulemuse.

Teoreem 9 *Olgu F mingi tõenäosusjaotuse jaotusfunktsioon ning olgu F^- selle üldistatud pöördfunktsioon. Siis juhusliku suuruse $X = F^-(Y)$, kus $Y \sim U(0,1)$, jaotusfunktsiooniks on F.*

Tõestus. Leiame juhusliku suuruse X jaotusfunktsiooni. Definitsiooni kohaselt
$$F_X(x) = P(\{X \leq x\}) = P(\{F^-(Y) \leq x\}),\ x \in \mathbf{R}.$$
Olgu elementaarsündmuste ruumiks Ω. Kuna lemma 8 kohaselt
$$\{\omega \in \Omega\ :\ F^-(Y(\omega)) \leq x\} = \{\omega \in \Omega\ :\ Y(\omega) \leq F(x)\},$$
siis
$$F_X(x) = P(\{Y \leq F(x)\}) = F_Y(F(x)).$$
Võttes arvesse, et ühtlase jaotusega $U(0,1)$ juhusliku suuruse Y jaotusfunktsiooni korral $F_Y(y) = y,\ 0 \leq y \leq 1$, siis olemegi näidanud, et
$$F_X(x) = F(x)\ \forall x \in \mathbf{R}.\square$$

Märkus: Tõenäosusteooria seisukohalt ei ole vahet, kas punkt valitakse juhulikult lõigust [0,1] või vahemikust (0,1), sest otspunktide saamise tõenäosus on esimesel juhul 0 ja seetõttu ei muutu tõenäosusarvutuse seisukohalt midagi, kui need võimalused ära jätta. Üldistatud pöördfunktsiooni kasutamisel on aga mugavam, kui ei tule väärtustega 0 ja 1 tegeleda.

4.1.2 Lõpliku arvu väärtustega diskreetse juhusliku suuruse esitamine ühtlase jaotuse kaudu

Vaatleme juhuslikku suurust, mille võimalikeks väärtusteks on $c_1 < c_2 \ldots < c_m$, olgu $p_1, p_2, \ldots, p_m$ vastavad tõenäosused. Sellise juhusliku suuruse jaotusfunktsioon on tükiti konstantne, katkevuspunktideks on $c_1, \ldots, c_m$ ja hüpete kõrguseks $p_1, \ldots, p_m$. Seega avaldub üldistatud pöördfunktsioon vahemikus $y \in (0, 1)$ kujul

$$F^-(y) = \begin{cases} c_1, & 0 < y \leq p_1, \\ c_2, & p_1 < y \leq p_1 + p_2 \\ \ldots & \\ c_m, & \sum_{i=1}^{m-1} p_i < y < 1. \end{cases}$$

Seega saame soovitud jaotusega juhusliku suuruse väärtusi genereerida, genereerides ühtlase jaotusega $U(0, 1)$ juhuslikke suuruseid ning rakendades neile eelneva valemi abil defineeritud üldistatud pöördfunktsiooni.

Peatükk 5

Loeng 5: Lihtne ja üldistatud valikumeetod

5.1 Valikumeetodid

Selle asemel, et kohe soovitud jaotusest arve genereerida, võime proovida seda tegevust jaotada lihtsamateks osadeks: genereerime punkte mõnest tuntud jaotusest ja siis "hõrendame"neid nii, et erinevatesse piirkondadesse jäävate punktide osakaalud vastaksid soovitud jaotusele. See ongi valikumeetodite idee.

5.1.1 Lihtne valikumeetod

Vaatleme pidevat juhuslikku suurust tihedusfunktsiooniga f. Eeldame, et f on 0 väljaspool lõiku [a,b] ning samuti eeldame, et f on ülalt tõkestatud konstandiga c, st $f(x) \leq c \ \forall x$. Valikumeetodi algoritmi intuitiivne kirjeldus on järgmine

1. Pommitame ristkülikut $[a, b] \times [0, c]$ selles juhuslikult valitud punktidega (vastavalt ühtlasele jaotusele)
2. Jätame alles ainult need, mis satuvad $y = f(x)$ alla X väärtusteks võtame saadud punktide x-koordinaadid

Osutub, et nii saame valimi tihedusfunktsioonile f vastavast jaotusest. Veelgi enam, algoritmi võib rakendada suvalise mittenegatiivse tõkestatud funktsiooni f korral ning tulemuseks saame valimi jaotusest, mille tihedusfunktsioon on proportsionaalne funktsiooniga f. Tõestame vastava tulemuse.

Teoreem 10 *Eeldame, et mittenegatiivne funktsioon f on 0 väljaspool lõiku [a,b] ning et f on ülalt tõkestatud konstandiga c, st $f(x) \leq c \ \forall x$. Olgu $Y_1 \sim U(a, b)$ ja $Y_2 \sim U(0, c)$ sõltumatud juhusliud suurused. Defineerime juhusliku suuruse X järgmiselt:*

$X = Y_1$ tingimusel, et $Y_2 \leq f(Y_1)$

Siis X on pidev juhuslik suurus tihedusfunktsiooniga, mis on proportsionaalne funktsiooniga f.

Tõestus. Leiame juhusliku suuruse X jaotusfunktsiooni avaldise. Definitsiooni kohaselt

$$F_X(x) = P(X \leq x) = P(Y_1 \leq x \mid Y_2 \leq f(Y_1)) = \frac{P(Y_1 \leq x,\, Y_2 \leq f(Y_1))}{P(Y_2 \leq f(Y_1)))}.$$

Tähistame

$$D_x = \{(y_1, y_2) \in \mathbb{R}^2 : y_1 \le x, y_2 \le f(y_1),$$

siis tingimus $\{Y_1 \le x,\, Y_2 \le f(Y_1)\}$ on esitatav kujul $(Y_1, Y2) \in D_x$. Teame, et

$$f_{Y_1}(y_1) = \begin{cases} \frac{1}{b-a}, & y_1 \in [a,b], \\ 0, & y_1 \notin [a,b], \end{cases} \quad f_{Y_2}(y_2) = \begin{cases} \frac{1}{c}, & y_2 \in [0,c], \\ 0, & y_2 \notin [0,c]. \end{cases}$$

Kuna Y_1 ja Y_2 sõltumatud, siis juhusliku vektori (Y_1, Y_2) tihedusfunktsioon avaldub kujul

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2),\ y_1, y_2 \in \mathbb{R}$$

ning pideva juhusliku vektori tihedusfunktsiooni omaduse 3 (vt lemma 27) kohaselt

$$P((Y_1, Y_2) \in D_x) = \int_{-\infty}^{x} \left(\int_{-\infty}^{f(y_1)} f_{Y_2}(y_2)\, dy_2 \right) f_{Y_1}(y_1)\, dy_1 = \int_{-\infty}^{x} \frac{f(y_1)}{c} f_{Y_1}(y_1)\, dy_1.$$

Arvestades, et $f_{Y_1}(y_1)$ on nullist erinev ainult piirkonnas $y_1 \in [a,b]$, tuleb eelneva integraali arvutamisel vaadelda kolme juhtu: $x < a$, $x \in [a,b]$, $x > b$. Seega saame

$$P((Y_1, Y_2) \in D_x) = \begin{cases} 0, & x < a, \\ \int_a^x \frac{f(y_1)}{c(b-a)}\, dy_1, & x \in [a,b], \\ \int_a^b \frac{f(y_1)}{c(b-a)}\, dy_1, & x > b. \end{cases}$$

Tähistades

$$c_f = \int_a^b f(y)\, dy,$$

saame

$$P(Y_2 \le f(Y_1)) = P((Y_1, Y_2) \in D_\infty) = \frac{c_f}{c(b-a)}$$

ning seega

$$F_X(x) = \begin{cases} 0, & x < a, \\ \int_a^x \frac{f(y)}{c_f}\, dy, & x \in [a,b], \\ 1, & x > b \end{cases}$$

Arvestades, et

$$f_X(x) = F_X'(x) = \begin{cases} \frac{f(x)}{c_f}, & x \in [a,b], \\ 0, & x \notin [a,b], \end{cases}$$

on teoreemi väide tõestatud.□

Märkus. Kasutades teadmist, et $U \sim U(0,1)$ korral $Y_1 = a + (b-a)U$ puhul $Y_1 \sim U(a,b)$ ja $Y_2 = cU$ puhul $Y_2 \sim U(0,c)$, saab valikumeetodit teostada, lähtudes jaotusele $U(0,1)$ vastavatest pseudojuhuslikest arvudest.

Valikumeetodi *efektiivsuseks* nimetatakse X defineerimisel kasutatud sündmuse toimumise tõenäosust, st lihtsa valikumeetodi puhul arvu

$$P(Y_2 \le f(Y_1)) = \frac{c_f}{c(b-a)}$$

5.1.2 Üldine valikumeetod

Lihtne valikumeetod võimaldab genereerida ainult tõkestatud väärtustega juhuslikke suuruseid. Suur osa huvipakkuvatest jaotustest on aga sellised, mille tihedusfunktsioon on nullist erinev kuitahes suurte (või väikeste) argumendi väärtuste korral, mistõttu ka vastava juhusliku suuruse väärtuste piirkond ei ole tõkestatud. Osutub, et valikumeetodit saab üldistada selliste juhtudega tegelemiseks

Teoreem 11 *Olgu f ja g mittenegatiivsed ning lõpliku graafikualust pindala omavad funktsioonid ja c selline reaalarv, et kehtib tingimus $f(x) \leq cg(x)\ \forall x$. Olgu Y_1 ja $Y2$ sõltumatud juhuslikud suurused, kusjuures Y_1 on tihedusfunktsioon on kujul*

$$f_{Y_1}(y_1) = \frac{g(y_1)}{c_g},\ c_g = \int_{-\infty}^{\infty} g(x)\,dx$$

ja Y_2 on ühtlase jaotusega $U(0,1)$. Defineerime juhusliku suuruse

$$X = Y_1 \mid \{g(Y_1) > 0, Y_2 \leq \frac{f(Y_1)}{c \cdot g(Y_1)}\}.$$

Siis juhusilk suurus X on funktsiooniga f proportsionaalsele tihedusfunktsioonile vastava jaotusega.

Tõestus. Tõestuse on analoogiline lihtsa valikumeetodi puhul toodud tõestusega, mistõttu selle läbikirjutamine jääb harjutuseks lugejale.□

Lihtne on veenduda, et üldise valikumeetodi efektiivsus avaldub kujul

$$P(Y_2 \leq \frac{f(Y_1)}{cg(Y_1)}) = \frac{c_f}{c \cdot c_g},$$

kus

$$c_f = \int_{-\infty}^{\infty} f(x)\,dx.$$

Peatükk 6

Loeng 6: Valikumeetodi eelduste kontrollist. Genereerimine segujaotusest. Jaotuste teisendamine. Genereerimine tinglikust jaotusest

6.1 Valikumeetodi eelduste kontrollist

Üldise valikumeetodi korral tuleb lähtejaotus valida nii, et selle tihedusega proportsionaalse funktsiooni g korral saab leida sellise positiivse kordaja c, et soovitud tihedus f rahuldab võrratust

$$f(x) \leq c\, g(x) \qquad -\infty < x < \infty.$$

Sellest tingimusest saab teha mitmesuguseid järeldusi.

- Kui mingi x korral $g(x) = 0$, siis $f(x)$ peab olema null. Teisipidi, kui $f(x) > 0$ mingi $x \in \mathbf{R}$ korral, siis peab kehtima $g(x) > 0$. Seega peab lähtejaotuse võimalike väärtuste piirkond sisaldama soovitud jaotuse kandjat.
- suhe $\frac{f(x)}{g(x)}$ peab olema tõkestatud piirkonnas $\{x \in \mathbf{R} : f(x) > 0\}$
- Juhul, kui hulk $f(x) > 0$ on tõkestamata, tuleb eriti hoolikalt kontrollida, mis juhtub piiril $x \to \infty$ või $x \to -\infty$. Sageli tuleb siin abiks teadmine, et $\lim_{x\to\infty} |Q(x)|e^{P(x)} = \infty$, kus Q on ratsionaalfunktsioon ja P on polünoom, kus suurima x astme kordaja on positiivne ning sama piirväärtus on null, kui P definitsiooni suurima x astme kordaja on negatiivne.
- Samuti on potentsiaalselt problemaatiline juhtum, kus piirkonnal $\{x \in \mathbf{R} : f(x) > 0\}$ on mingi rajapunkt $x_0 \in \mathbf{R}$, millele lähenedes nii f kui g lähenevad nullile või on mõlemad tõkestamata. Ka sel juhul tuleb uurida matemaatlisest analüüsist (või kõrgemast matemaatikast) tuttavate võtetega piirväärtuse $\lim_{x\to x_0, f(x)>0} \frac{f(x)}{g(x)}$ käitumist.

Mõningad jaotused, mida saab sageli kasutada lähtejaotusena või mille abil saab lihtsalt defineerida sobivaid lähtejaotuseid, on järgmised:

- Normaaljaotus $N(\mu,\sigma)$ - tihedus $f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ on positiivne kõikjal, suurte $|x|$ väärtuste korral käitub tihedus nagu $e^{-\frac{x^2}{2\sigma^2}}$
- Eksponentjaotus $Exp(\lambda)$ - tihedus $f_X(x)=\lambda e^{-\lambda x}I_{x>0}(x)$ on positiivne piirkonnas $x\geq 0$, suure x väärtuse korral käitumine $e^{-\lambda x}$
- Cauchy jaotus $Cauchy(x_0,\gamma)$ - tihedus $f_X(x)=\frac{1}{\pi\gamma(1+\frac{(x-x_0)^2}{\gamma})}$ on kõikjal positiivne, suurte $|x|$ väärtuste korral käitub nagu $\frac{1}{x^2}$

Järgnevalt vaatleme veel võimalusi tuntud jaotuste abil soovitud jaotusest väärtuste genereerimise võimalusi.

6.2 Genereerimine segujaotusest

Sageli võib andmete jaotusi uurides täheldada mitut "küüru"või lihtsalt erinevat tüüpi käitumist erinevates väärtuste piirkondades, mistõttu ei vasta nende käitumine lihtsatele tuntud jaotustele. Tihti on selline käitumine põhjustatud üldkogumi mittehomogeensusest ehk sellest, et üldkogumis on mitu erinevat rühma (näiteks mehed/naised), milles mõõdetav tunnus käitub erinevalt. Olgu meil m rühma, kusjuures igas rühmas käitugu uuritav tunnus vastavalt pidevale juhuslikule suurusele tihedusfunktsiooniga $f_i,\ i=1,2,\ldots,m$. Olgu iga rühma osakaal üldkogumis p_i, siis täistõenäosuse valemi (vaata lemma 22) abil on lihtne näidata, et uuritava tunnuse tihedusfunktsioon üldkogumist juhusliku liikme valikul avaldub kujul

$$f(x)=\sum_{i=1}^{m}p_if_i(x),\ -\infty<x<\infty.$$

Tõestame selle valemi.

Teoreem 12 *Olgu üldkogumis m rühma, kusjuures igas rühmas käitugu uuritav tunnus vastavalt pidevale juhuslikule suurusele tihedusfunktsiooniga $f_i,\ i=1,2,\ldots,m$. Siis uuritava tunnuse tihedusfunktsioon üldkogumist juhusliku liikme valikul avaldub kujul*

$$f(x)=\sum_{i=1}^{m}p_if_i(x),\ -\infty<x<\infty.$$

Tõestus Olgu X uuritava tunnuse väärtus üldkogumist juhuslikult valitud liikme korral ning olgu A_i sündmus, et valitud liige kuulub rühma i. Tähistagu F_i tihedusfunktsioonile f_i vastavat jaotusfunktsiooni. Täistõenäosuse valemi kohaselt

$$F_X(x)=P(X\leq x)=\sum_{i=1}^{m}P(A_i)P(X\leq x\mid A_i).$$

Ülesande teksti põhjal on juhusliku suuruse X tinglik jaotus tingimusel, et valituks osutub rühma A_i liige, antud jaotusfunktsioniga F_i, seega

$$F_X(x)=\sum_{i=1}^{m}p_iF_i(x).$$

Kuna tihedusfunktsioon on jaotusfunktsiooni tuletis, saame nüüd

$$f_X(x) = \sum_{i=1}^{m} p_i F_i'(x) = \sum_{i=1}^{m} p_i f_i(x),\ -\infty < x < \infty.$$

Teoreem on tõestatud.□

Siit saame algoritmi etteantud tihedusfunktsiooniga juhusliku suuruse väärtuste genereerimiseks segujaotuse meetodil:

1. Leiame etteantud tihedusfunktsiooni f esituse tuntud tihedusfunktsioonide f_i kaudu kujul
$$f(x) = \sum_{i=1}^{m} p_i f_i(x),\ x \in \mathbb{R}$$
2. Genereerime i väärtuse jaotusest $(j, p_j),\ j = 1, \ldots, m$
3. Genereerime X väärtuse tihedusele f_i vastavast jaotusest.

6.3 Pideva juhusliku suuruse teisendused

Mõnikord on võimalik soovitud jaotusega juhuslikku suurust lihtsalt saada nii, et rakendame tuntud jaotusega juhuslikule suurusele mingit funktsiooni. Rangelt monotoonsete funktsioonide rakendamisel on võimalik anda lihtsa seose esialgse ja teisendatud juhusliku suuruse tihedusfunktsioonide vahel.

Lemma 13 *Kui X on pidev juhuslik suurus tihedusfunktsiooniga f ja g on rangelt monotoonne funktsioon pöördfunktsiooniga h (st $h(g(x)) = x\ \forall x$), siis juhusliku suuruse $Y = g(X)$ tihedusfunktsioon avaldub kujul*

$$f_Y(y) = \begin{cases} f_x(h(y))|h'(y)|, & y \in g(R) \\ 0, & y \notin g(R) \end{cases}$$

Tõestus. Tõestus on harjutuseks lugejale. Soovitus: kõigepealt saab leida seose jaotusfunktsioonide vahel ja seejärel tuletise abil seose tihedusfunktsioonide vahel. Eraldi tuleb käsitleda rangelt kasvavat funktsiooni g ja rangelt kahanevat funktsiooni g, sest võrratusele rangelt kahaneva funktsiooni rakendamisel saame esialgsega vastupidise võrratuse.□

6.4 Genereerimine tinglikust jaotusest

Vaatleme juhtu, kus me tahame genereerida juhusliku suuruse väärtuseid tingimusel, et need väärtused jäävad teatud vahemikku. Leiame juhusliku suuruse $Y = X \mid \{a < X \leq b\}$ jaotusfunktsiooni:

$$F_Y(y) = P(Y \leq y) = P(X \leq y \mid a < X \leq b) = \frac{X \leq y, a < X \leq b}{P(a < x \leq b)}.$$

Kuna

$$P(a < x \leq b) = F_X(b) - F_X(a)$$

ning

$$\{X \le y, a < X \le b\} = \begin{cases} \emptyset, & y \le a, \\ \{a < X \le y\}, & y \in (a, b], , \\ \{a < X \le b\} & y > b \end{cases}$$

siis

$$F_Y(y) = \begin{cases} 0, & y \le a, \\ \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)}, & a < y \le b, \\ 1, & y > b. \end{cases}$$

Kui X on pidev juhuslik suurus, siis siit järeldub, et tingliku jaotuse tihedusfunktsioon avaldub kujul

$$f_Y(y) = \begin{cases} 0, & y \le a, \\ \frac{f_X(y)}{F_X(b) - F_X(a)}, & a < y \le b, \\ 0, & y > b. \end{cases}$$

Osutub, et tinglikust jaotusest on lihtne väärtuseid genereerida, kasutades jaotusfunktsiooni pöördfunktsiooni.

Teoreem 14 *Olgu X juhuslik suurus jaotusfunktsiooniga F_X. Olgu U juhuslik suurus jaotusega $U(F_X(a), F_X(b))$, kus $a < b$ on mingid reaalarvud. Siis juhuslik suurus $Y = F_X^-(U)$, kus F_X^- on funktsiooni F_X üldistatud pöördfunktsioon, on sama jaotusega kui $X \mid \{a < X \le b\}$*

Tõestus. Leiame juhusliku suuruse Y jaotusfunktsiooni. Lemmat 8 kasutades saame

$$F_Y(y) = P(F_X^-(U) \le y) = P(U \le F_X(y)).$$

Ühtlase jaotuse jaotusfunktsiooni avaldist kasutades saame nüüd

$$F_Y(y) = \begin{cases} 0, & F_X(y) \le F_X(a), \\ \frac{F_X(y) - F_X(a)}{F_X(y) - F_X(a)}, & F_X(a) < F_X(y) \le F_X(b), \\ 1, & F_X(y) > F_X(b). \end{cases}$$

Arvestades funktsiooni F_X monotoonset kasvamist, on lihtne veenduda, et eelnev valem annab iga y korral sama väärtuse, mis tingliku jaotuse $X \mid \{a < X \le b\}$ jaotusfunktsiooni valem, seega on teoreem tõestatud. □

Peatükk 7

Loeng 7: Pseudojuhuslike vektorite genereerimine. Pidevate juhuslike vektorite teisendused

Pidevate juhuslike vektorite pööratavate teisenduste puhul kehtib järgmine üldine seos tihedusfunktsioonide vahel.

Teoreem 15 *Olgu X pidev juhuslik vektor, mis võtab väärtusi piirkonnas $G \subset R^k$ ja mille tihedusfunktsioon on f_X. Olgu $g = (g_1, g_2, \ldots, g_k) : G \to H$, $H \subset R^k$ pööratav teisendus, mille kõik komponendid g_i on diferentseeruvad ja mille pöördteisendus on $g^{-1} = (h_1, h_2, \ldots, h_k)$. Sellisel juhul juhusliku vektori $Y = g(X)$ tihedusfunktsioon on*

$$f_Y(y) = f_X(g^{-1}(y))\ abs\left(\begin{vmatrix} \frac{\partial h_1(y)}{\partial y_1} & \frac{\partial h_1(y)}{\partial y_2} & \cdots & \frac{\partial h_1(y)}{\partial y_k} \\ \frac{\partial h_2(y)}{\partial y_1} & \frac{\partial h_2(y)}{\partial y_2} & \cdots & \frac{\partial h_2(y)}{\partial y_k} \\ \vdots & \vdots & & \vdots \\ \frac{\partial h_k(y)}{\partial y_1} & \frac{\partial h_k(y)}{\partial y_2} & \cdots & \frac{\partial h_k(y)}{\partial y_k} \end{vmatrix}\right),\ kui\ y \in H.$$

Absoluutväärtus maatriksist tähistab eelnevas teoreemis selle maatriksi determinanti.

7.0.1 Tihedusfunktsiooni teisenemine lineaarsete teisenduste korral

Olgu X pidev m-mõõtmelinejuhuslik vektor ja f_X selle tihedusfunktsioon. Vaatleme dimensiooniga $m \times m$ pööratava maatriksi A ja vektori $b = (b_1, \ldots, b_m)^T$ poolt defineeritud teisenduse

$$g(x) = Ax + b$$

rakendamisel saadavat juhuslikku suurust $Y = AX + b$. Kuna pöördteisendus avaldub sel juhul kujul

$$h(y) = A^{-1}(y - b),$$

saame teoreemi 15 põhjal avaldada Y tihedusfunktsiooni kujul

$$f_Y(y) = f_X(A^{-1}(y - b))\,\mathrm{abs}(|A^{-1}|).$$

7.1 Normaaljaotusega juhuslike suuruste ja juhuslike vektorite genereerimine

Osutub, et mõnikord on lihtsam genereerida sõltumatute juhuslike suuruste paare kui üksikuid väärtuseid. Üks näide selle kohta on Box-Mülleri meetod sõltumatute standardsete normaaljaotusega juhuslike suuruste genereerimiseks.

7.1.1 Box-Mülleri meetod

Meetod põhineb teadmisel, et standardse kahemõõtmelise normaaljaotusega vektori komponendid on sõltumatud normaaljaotusega juhuslikud suurused. Algoritm on järgmine:

1. genereeri kaks pseudojuhuslikku arvu u_1, u_2 jaotusest $U(0,\,1)$,
2. arvuta $r = \sqrt{-2 \cdot \ln(u_1)},\ \theta = 2\pi \cdot u_2$,
3. arvuta $x_1 = r \cdot \cos(\theta)$, $x_2 = r \cdot \sin(\theta)$.

Saadud väärtused x_1 ja x_2 vastavad sõltumatute jaotusega $N(0,1)$ juhuslike suuruste väärtustele.
Veendume eelnevas väites, kasutades teoreemi 15. Alustame sellest , et $U = (U_1, U_2)$ on ühtlase jaotusega ühikruudul, seega

$$f_U(u_1, u_2) = \begin{cases} 1, & 0 < u_1 < 1,\, 0 < u_2 < 1, \\ 0 & \text{mujal.} \end{cases}$$

Box-Mülleri meetod vastab sellele vektorile teisenduse

$$g(u_1, u_2) = (\sqrt{-2 \cdot \ln(u_1)} \cdot \cos(2\,\pi u_2), \sqrt{-2 \cdot \ln(u_1)} \cdot \sin(2\,\pi u_2))$$

rakendamises. Teisendatud juhusliku suuruse tihedusfunktsiooni avaldise leidmiseks läheb meil vaja pöördteisenduse valemeid. Pöördteisenduse esimene funktsioon on lihtsalt leitav:

$$h_1(x_1, x_2) = e^{-\frac{x_1^2 + x_2^2}{2}}$$

Pöördteisenduse teise funktsiooni puhul tuleb arvestada, et funktsioon arctan väärtused on piirkonnas $(-\frac{\pi}{2}), \frac{\pi}{2}$, seega tuleb olla hoolikas erinevatele tasandi veeranditele vastavate punktide teisendamisel. Tulemuseks on:

$$h_2(x_1, x_2) = \frac{1}{2\pi} \begin{cases} \arctan(\frac{x_2}{x_1}), & x_1 > 0, x_2 > 0, \\ \frac{\pi}{2}, & x_1 = 0, x_2 > 0, \\ \pi + \arctan(\frac{x_2}{x_1}), & x_1 < 0, \\ \frac{3\pi}{2}, & x_1 = 0, x_2 < 0, \\ 2\pi + \arctan(\frac{x_2}{x_1}), & x_1 > 0, x_2 < 0. \end{cases}$$

Edasine kontroll (jakobiaani leidmine ja veendumine, et teisendatud vektori tihedusfunktsioon on tõepoolest standardse kahemõõtmelise normaaljaotuse tihedusfunktsioon) on harjutus lugejale.

7.1.2 Jaotusega $N(\mu, \Sigma)$ juhuslike vektorite genereerimine

Soovime genereerida juhuslikke vektoreid jaotusega $N(\mu, \Sigma)$, kus μ on m-mõõtmeline vektor ja Σ on pööratav $m \times m$ kovariatsioonimaatriks. Vaadeldes lineaarteisenduse $Y = AX + \mu$ abil saadud juhusliku vektori tihedusfunktsiooni, on lihtne veenduda, et standardse normaaljaotuse teisendamisel saame normaaljaotuse, mille keskväärtus on μ ja kovariatsioonimaatriks on $\Sigma = A\,A^T$. Seega tuleb etteantud Σ korral leida selline A, et $\Sigma = A\,A^T$. Üks võimalus selleks on defineerida maatriksi V, mille veergudeks on Σ normeeritud omavektorid ja diagonaalmaatriksi Λ, kus diagonaalil on Σ omaväärtused. Siis maatriksi $A = V\Lambda^{\frac{1}{2}}$ korral kehtib $A\,A^T = \Sigma$ (veendu selles!)

Peatükk 8

Loeng 8: Juhuslike vektorite genereerimine tinglike jaotuste abil. Valikumeetodid mitmemõõtmelisel juhul. Metropolis-Hastings meetod

8.1 Juhuslike vektorite genereerimine tinglike jaotuste abil

Teame tõenäosuse korrutamise reeglit

$$P(A_1A_2\cdots A_m) = P(A_1)P(A_2 \mid A_1) \cdots P(A_m \mid A_1\cdots A_{m-1})$$

Osutub, et sarnane valem kehtib ka pideva juhusliku vektori tihedusfunktsiooni jaoks

$$f_X(x_1,\ldots,x_m) = f_{X_1}(x_1)f_{X_2|X_1}(x_2 \mid x_1) \cdots f_{X_m|X_1,X_2,\ldots,X_{m-1}}(x_m \mid x_1,x_2,\ldots,x_{m-1}),$$

kus tinglikud tihedusfunktsioonid saab leida samm-sammult seostest

$$f_{X_1}(x_1)f_{X_2|X_1}(x_2 \mid x_1)\cdots f_{X_k|X_1,X_2,\ldots,X_{k-1}}(x_k \mid x_1,x_2,\ldots,x_{k-1}) = \underbrace{\int\cdots\int}_{m-k} f(x_1,\ldots,x_m)dx_{k+1}\cdots dx_m,\ 1\le k\le m,$$

Siin 0-kordne integraal tähistab lihtsalt integraalialust funktsiooni. Seega, kui suudame nendele tinglikele tihedustele vastavaid juhuslikke suuruseid genereerida, saame soovitud jaotusega juhusliku vektori tekitada samm-sammult: kõigepealt genereerime X_1 väärtuse vastavalt tihedusele f_{X_1}, seejärel genereerime X_2 vastavalt tihedusele $f_{X_2|X_1}$ jne.

Näide. Vaatleme kolmnurgas $D = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le 2x\}$ ühtlase jaotusega vektori (X,Y) genereerimist. Paneme tähele, et D pindala on 1. Leiame suuruse X tihedusfunktsiooni. Selleks peame leidma integraali

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

Arvestades, et ühtlase jaotuse tihedus on konstantne väärtusega $\frac{1}{|D|}$vaadeldavas piirkonnas ning null väljaspool, saame

$$f_X(x) = \begin{cases} \int_0^{2x} 1\,dy = 2x, & 0 \le x \le 1, \\ 0 & \text{mujal.} \end{cases}$$

Tinglik tihedus $f_{Y|X}$ on leitav valemist

$$f_X(x) f_{Y|X}(y \mid x) = f_{X,Y}(x, y),$$

kust järeldub, et $Y|X$ on ühtlase jaotusega piirkonnas $[0,\ X]$ (tihedus on $\frac{1}{2x}$, kui $0 < y < 2x$ ja null mujal). Seega saab sobivaid vektoreid, kasutades näiteks jaotusfunktsiooni pööramise meetodit X genereerimiseks ning seejärel eelkirjeldatud ühtlasest jaotusest genereerimist vastava Y väärtuse saamiseks.

8.2 Lihtne ja üldine valikumeetod mitmemõõtmelisel juhul

Valikumeetodid töötavad mitmemõõtmelisel juhul sarnaselt ühemõõtmelisele juhule - matemaatiliselt tuginevad need ikka sellele, et kui valime ühtlase jaotusega m-mõõtmelisi punkte mingi funktsiooni graafiku alusest piirkonnast, siis nende punktide esimesed $m-1$ koordinaati käituvad graafiku tekitamise funktsioonile vastava tihedusfunktsiooniga jaotusele.

8.2.1 Lihtne valikumeetod

Eeldame, et soovitud jaotuse tihedusfunktsiooniga võrdeline funktsioon $f_{X,Y}$ on nullist erinev ainult ristkülikus $[a_1,\ b_1] \times [a_2,\ b_2]$ ning et see on tõkestatud konstandiga c.

Lihtsa valikumeetodi algoritm on järgmine:

1. Tekitame n pseudojuhuslikku arvu y_1 jaotusest $U(a_1, b_1)$, y_2 jaotusest $U(a_2, b_2)$ ja y_3 jaotusest $U(0, c)$
2. Jätame alles need paaridest (y_1, y_2), mille korral kehtib $y_3 \le f_{X,Y}(y_1, y_2)$

8.2.2 Üldistatud valikumeetod

Eeldame, et oskame genereerida vektorit (Y_1, Y_2), millele vastav tihedusfunktsioon g_{Y1,Y_2} rahuldab mingi $c > 0$ korral tingimust

$$f_{X,Y}(x, y) \le c \cdot g_{Y1,Y_2}(x, y)\ \forall (x, y) \in \mathbb{R}^2$$

Algoritm on järgmine

1. Tekitame n pseudojuhuslikku vektorit (y_1, y_2) tihedusfunktsioonile g_{Y_1,Y_2} vastavast jaotusest ning n pseudojuhuslikku arvu y_3 jaotusest $U(0, 1)$
2. Jätame alles need paaridest (y_1, y_2), mille korral kehtib $y_3 \le \frac{f_{X,Y}(y_1,y_2)}{c \cdot g_{Y1,Y_2}(y_1,y_2)}$

Sobivat mitmemõõtmelist lähtejaotust on sageli võimalik leida sõltumatute komponentidega jaotuste hulgast.

8.3 Metropolis-Hastings algoritm

Mitmemõõtmelise jaotuse puhul sageli väga raske leida tinglikke jaotuseid, samuti on raske valida head lähtejaotust valikumeetodi jaoks. Samas juhuslike protsesside puhul sageli kehtib tulemus, et pärast küllalt pikka ajavahemikku on protsessi mingile konkreetsele ajamomendile vastav väärtus teatud kindla jaotusega (nn statsionaarse jaotusega) juhuslik suurus sõltumata sellest, milline oli protsessi algväärtus. Siit tuleb idee vaadelda protsessi, kus m-mõõtmeline punkt eksleb juhuslikult nii, et liikumine soodustab sattumist sinna, kuhu soovitud tõenäosusega juhusliku vektori väärtused peaks sattuma suurema tõenäousega ja harvemini satutakse sinna, kuhu vaadeldava vektori väärtused satuvad väikese tõenäosusega. Osutub, et liikumist saab organiseerida nii, et statsionaarseks jaotuseks on meie soovitud jaotus.

Vaatleme algoritmi erijuhtu, kus ekslemisel uue asukoha määramisel kasutame normaaljaotust (nn Metropolise algoritm). Eeldame, et teame soovitud tihedusfunktsiooniga proportsionaalset funktsiooni f. Algoritm on järgmine:

1. Valime algpunkti x_1 (NB! peab olems selline, et $f(x_1) > 0$)
2. iga $i = 1, 2, ..., n-1$ korral
   (a) leiame uue asukoha kandidaadi $x \sim N(x_i, \Sigma)$
   (b) genereerime $u \sim U(0, 1)$
   (c) Kui $u \leq \frac{f(x)}{f(x_i)}$, siis $x_{i+1} = x$, muidu $x_{i+1} = x_i$
3. Väljastame punktid x_{n_0}, $x_{n_0+\ell}$, $x_{n_0+2\ell}$ jne, kus n_0 ja ℓ on sobivalt valitud naturaalarvud

Algoritmi puhul on oluline mõista, et kuigi pärast piisavalt pikka aega n_0 vastavad väärtused x_i soovitud jaotusega juhuslike suuruste väärtustele, ei vasta järjestikused x_i väärtused sõltumatute katsete tulemustele. Sellest on tingitud täiendava parameetri ℓ valik, mis peab olema piisavalt suur, et sõltuvus eelmisest valitud väärtusest oleks praktiliselt olematu.

Peatükk 9

Loeng 9: Juhuslike vektorite genereerimine Gibbsi valiku meetodil

9.1 Gibbsi valik

Mõnikord on raske leida tinglikke jaotuseid juhul, kui üle ühe vektori koordinaatidest on fikseeritud, kuid ainult ühe koordinaadi fikseerimise korral on lihtne kindlaks teha, milline on vaadeldavale jaotusele vastav tinglik jaotus. Sel juhul on võimalik kasutada Gibbsi valiku algoritmi.

Eeldame, et suudame genereerida väärtuseid tinglikest jaotustest

$$X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m,\ i = 1, \ldots, m$$

Algorit on järgmine:

1. Valime algpunkti $(x_{11}, x_{12}, \ldots, x_{1m})$
2. iga $i = 2, 3, \ldots, n$ korral genereerime
   (a) x_{i1} jaotusest $X_1 \mid X_2 = x_{i-1,2}, \ldots, X_m = x_{i-1,m}$,
   (b) x_{i2} jaotusest $X_2 \mid X_1 = x_{i,1}, X_3 = x_{i-1,3}, \ldots, X_m = x_{i-1,m}$,
   (c) x_{i3} jaotusest $X_3 \mid X_1 = x_{i,1}, X_2 = x_{i,2}, X_4 = x_{i-1,4}, \ldots, X_m = x_{i-1,m}$,
   (d) ...
   (e) x_{im} jaotusest $X_m \mid X_1 = x_{i,1}, X_2 = x_{i,2}, \ldots, X_{m-1} = x_{i,m-1}$,

Jällegi tuleb arvestada, et järjestikused punktid ei vasta enamasti sõltumatutele juhuslikele vektoritele, seetõttu tuleks kasutada neist ainult piisavalt hõredalt (indeksi i mõttes) valitud punkte!

Vaatleme mõningaid näiteid.

Näide 16 *Olgu inimsese ülekaalulisuse (X, 0-ei, 1-jah) ja ülemise vererõhu (Y) korral teada ,et $X \mid Y \sim Be(p(Y))$ ja $Y \mid X \sim N(\mu(X), 15)$, kus $p(y) = \frac{1}{1-e^{-0.1\cdot(y-140)}}$ ja $\mu(x) = 120 + 15 \cdot x$. Eesmärgiks genereerida Gibbsi meetodiga (enam-vähem) sõltumatuid katsetulemusi suuruste X ja Y ühisjaotusest.*

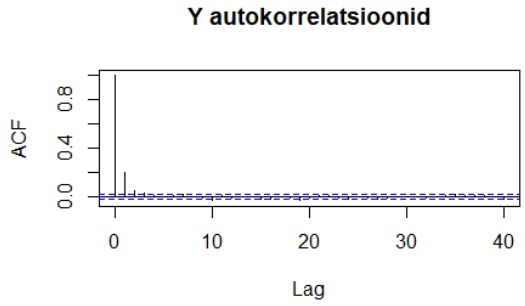
Selleks

- *Valime sobiva alguspunkti, näiteks* $(x_0 = 0, y_0 = 125)$.

- *Iga* $i = 1, 2, \ldots, n$ *korral genereerime*

  1. x_i *jaotusest* $Be(p(y_{i-1})$
  2. y_i *jaotusest* $N(\mu(x_i), 15)$.

- *Uurime, millise sammuga tuleks katsetulemusi valida, et ligikaudset sõltumatust saavutada. Praeguses näites võib näha sarnast pilti (R funktsiooniga* `acf()`*):*



Nagu näha, ei ole järjestikused punktid kindlasti sõltumatud, aga juba paari punkti vahelejätmisel kaovad vähemalt selgelt nullist erinevad autokorrelatsioonid ära. Sarnane pilt avaneb ka X *autokorrelatsioone vaadates. Seega võib paljude simuleerimisülesannete jaoks sobivate väärtustepaaride saamiseks väljastada tekitatud punktijadast näiteks iga viienda.*

Peatükk 10

Loeng 10: MC meetodite rakendused: p-väärtuste arvutamine ja võimsuse leidmine

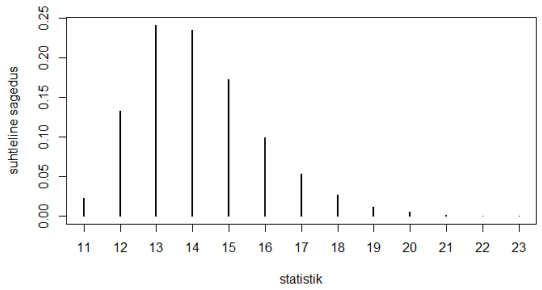
10.1 Statististilise testi p-väärtuse leidmine

Küllalt harva on kasutatava teststatistiku jaotus nullhüpoteesi kehtimisel täpselt teada iga valimimahu korral. Sageli on siiski teada selle asümptootiline jaotus ja seega on võimalik p-väärtust arvutada eeldusel, et valim on piisavalt suur. Kui aga on võimalik nullhüpoteesile vastavaid valimeid genereerida, siis saab simulatsioonide abil leida ligikaudselt p-väärtuseid ka juhtudel, kui statistiku piirjaotus ei ole teada või kui on kahtluseid, kas valimimaht on piirjaotuse kasutamiseks piisavalt suur.

Testi p-väärtuse leidmise eeskiri simulatsioonide kasutamisel on järgmine:

1. genereerime huvipakkuva valimimahu korral valimeid nullhüpoteesi kehtimise eeldusel;
2. leiame iga genereeritud valimi korral statistiku väärtuse;
3. kasutame saadud valimit statistiku väärtustest selleks, et leida ligikaudselt tõenäosus, et statistiku väärtus on vähemalt sama suur, kui huvipakkuva valimi korral leitud väärtus.

Näide 17 *Vaatleme täringuviskeid ning olgu statistikuks kõige sagedamine esinenud silmade arvu esinemiste arv katseseerias. Oletame, et vaadeldavas 60-st viskest koosnevas katseseerias tuli kõige rohkem kolmesid (18 korda). Eesmärgiks on leida selle statistiku väärtuse p-väärtus. Selleks teostame 10000 simulatsiooni 60-st viskest koosneva katseseerijast ausa täringu korral ja leiame iga vastavad statistiku väärtused. Saadud väärtuste sagedustabel on järgmine:*



Siit on näha, et statistiku väärtus võib ausa tärinug korral olla vähemalt 18 suhteliselt väikese tõenäosusega, täpsemalt tule vaadeldavas simulatsioonis selle tõenäosuse (ehk väärtuse 18 p-väärtuse) hinnanguks $0,046$.

10.2 Statististilise testi võimsuse leidmine

Statistilise testi rakendamisel saame me valida esimest liiki vea (H_1 vastuvõtmise, kui H_0 kehtib) tõenäosust, kuid sageli huvitab aga ka teist liiki viga - kui suur on võimalus, et me jääme nullhüpoteesi juurde siis, kui H_0 tegelikult ei kehti. Tuleb aga silmas pidada, et teist liiki vea tõenäosust saab arvutada ainult konkreetse eelduste tegemisel H_1-le vastava jaotuse kohta. Sageli pakub rakendustes huvi keskväärtuse erinemine H_0 jaotusele vastavast keskväärtusest teatud suuruse võrra, st H_1 kehtimisel eeldatakse, et vaatlused pärinevad jaotusest, mis erineb H_0-st ainult uuritava tunnuse keskväärtuse põhjal.

Tihti kasutatakse võimsuse hindamist selleks, et määrata valimi suurust, mis on vajalik soovitud suurusega efekti (keskväärtuse muutumise) avastamiseks etteantud tõenäosusega. See on suure praktilise tähtsusega näiteks juhul, kui katsete tegemine on kulukas.

Näide 18 *Leiame Kolmogorov-Smirnovi testi võimsuse valimisuurus 100 ja* $\alpha = 0.05$ *korral juhul, kui nullhüpoteesiks on andmete vatamine jaotusele* $Exp(1)$*, aga alternatiiviks on vastamine jaotusele* $Exp(0.8)$*. Selleks:*

1. *genereerime* $m = 1000$ *valimit suurusega* $n = 100$ *jaotusest* $Exp(0.8)$*;*
2. *rakendame iga valimi korral KS testi jaotusele* $Exp(1)$ *vastavuse kontrollimiseks*
3. *leiame kordade arvu, mille puhul KS testi p-väärtus on väiksem kui* α *ning jagame selle katsete arvuga* m*.*

Tulemus põhjal saab öelda, et võimsus on ligikaudu 0.4 (tulemus on juhuslik, saadud hinnanu 95% tõenäosusega kehtiv veahinnang on ligikaudu 0.01).

Peatükk 11

Loeng 11: Imputeerimine

Sageli on andmetes puuduvaid väärtuseid, mis muudavad analüüsi teostamise raskeks. Seega on loomulik idee lüngad enne analüüsimist mingite väärtustega täita ning seda protsessi nimetatakse imputeerimiseks. Samas on mitmete imputeerimise meetodite korral oht saada hilisemas andmeanalüüsis valesid tulemusi. Et võimalikke probleeme paremini kirjeldada, vaatleme järgnevalt erinevaid puuduvate andmete tekkimise iseloomu kirjeldavaid juhte.

11.1 Puudumise tekkimise mehhanismid

Puuduvate andmete korral on väga tähtis järele mõelda, millest puudumine võib olla mõjutatud. Imputeerimise kirjanduses eristatakse tavapäraselt järgmisi mehhanisme:

1. MCAR (*Missin Completely At Random*) - mingi väärtuse puudumine ei sõltu kuidagi kogutavatest andmetest;
2. MAR (*Missing At Random*) - puudumine ei sõltu konkreetse andmeelemendi väärtusest;
3. MNAR (*Missing Not At Random*) - puudumine ei sõltub puuduvatest väärtustest

Viimasel juhul on andmeanalüüs oluliselt raskendatud ja nõuab puudumiste tekke modelleerimist. Seetõttu piirdume valdavalt esimese kahe juhu vaatlemisega. Eelnevad sõnalised kirjeldused on siiski veidi ebatäpsed, seetõttu toome vastavad matemaatilised definitsioonid. Eelnevalt on meil vaja aga mõningaid tähistusi:

- Y - $n \times p$ maatriks p tunnuse väärtustest (ka nendest, mis andmestikus puuduvad);
- R - $n \times p$ 0–1 maatriks olemasolu indikaatori väärtustest;
- Kui y_{ij} väärtus on andmetes olemas, siis $r_{ij} = 1$; kui y_{ij} on puudu, siis $r_{ij} = 0$;
- Y_{obs} - maatriksi Y väärtused, mis andmestikus on olemas;
- Y_{mis} - Y väärtused, mis andmestikus puuduvad.

Seega puuduvate väärtustega andmestikku saab vaadelda juhuslike maatriksite paari $(\mathbf{Y}, \mathbf{R})$ konkreetse väärtusena. Eelnevaid tähistusi kasutades saab MCAR ja MAR tingimused panna kirja tõenäosusteooria keeles järgmiselt:

- MCAR tingimus:
$$P(\mathbf{R} = R | Y_{\text{obs}}, Y_{\text{mis}}) = P(\mathbf{R} = R)$$

- MAR tingimus
$$P(\mathbf{R} = R|Y_{\text{obs}}, Y_{\text{mis}}) = P(\mathbf{R} = R|Y_{\text{obs}})$$

Järgnevalt vaatleme meetodeid, kuidas puuduvate andmete probleemiga tegeletakse

11.2 Naiivsed meetodid

Kõige lihtsamini kasutatavad ideed puuduvate väärtustega tegelemiseks on järgmised.

- **Täielike vaatluste analüüs** - puuduvate väärtustega vaatlused eemaldatakse andmestikust. MCAR puhul tekib nii analüüsiks sobiv valim, kuid suure arvu eemaldatavate andmeridade korral võib palju infot kaduma minna. MAR puhul aga on suur oht saada valimit, mille põhjal saadud hinnangud on nihkega ja tehtavad järeldused huvipakkuvate hüpoteeside kohta ebakorrektsed.
- **Aritmeetilise keskmise/mediaani/moodiga asendamine** - iga tunnuse puuduvad väärtused asendatakse vastava arvkarakteristikuga. MCAR kehtimisel jäävad tunnuste põhjal arvutatud vastavad arvkarakteristikute hinnangud korrektseks, kuid tunnustevahelised statistilised seosed muutuvad nõrgemaks ja dispersiooni hinnangud muutuvad petlikult väikeseks. MAR puhul on lisaks võimalik hinnangute nihete suurenemine.
- **Mudelipõhine imputeerimine** - sobitatakse mingi statistiline mudel puuduvate väärtustega vaatluse prognoosimiseks teiste tunnuste abil ja täidetakse lüngad mudeli prognoosidega. Sobiva mudeli kasutamisel võivad ka MAR andmete puhul hinnangute nihked oluliselt väheneda, kuid probleemiks on tunnustevaheliste seoste võimendamine ja parameetrite hinnangute dispersiooni hinnangud ei tule usaldusväärsed.

Seega on sellised meetodid suurte puudustega ning pigem ei tohiks neid praktilises andmeanalüüsis enamikel juhtudel kasutada.

11.3 Mitmene imputeerimine

Kui meil on võimalik puuduvaid väärtuseid asendada nii, et need on valitud juhuslikut õigest tinglikust jaotuses (tingimusel, et vaadeldud väärtused on need, mis algses andmestikus), siis on võimalik leida huvipakkuvatele üldkogumi parameetritele hinnangud järgmiste nn Rubini reeglite abil. Selleks teeme järgmised eeldused:

- hindame mingit üldkogumi parameetrit θ;
- täieliku valimi korral on teada hinnangufunktsioon $\hat{\theta}(Y)$ ja hinnang selle dispersioonile $\hat{D}(Y)$.

Olgu $Y^{(i)}$ imputeeritud valim, $i = 1, 2, \ldots, m$. Siis Rubini reeglid on järgmised:

- Parameetri hinnang:
$$\hat{\theta}(Y_{\text{obs}}) = \frac{\sum_{i=1}^{m} \hat{\theta}(Y^{(i)})}{m}$$
- Dispersiooni hinnang:
$$\hat{D}(Y_{\text{obs}}) = \sum_{i=1}^{m} \frac{\sum_{i=1}^{m} \hat{D}(Y^{(i)})}{m} + (1 + \frac{1}{m}) \frac{\sum_{i=1}^{m} (\hat{\theta}(Y^{(i)}) - \hat{\theta}(Y_{\text{obs}}))^2}{m - 1}$$

Peatükk 12

Loeng 12: bootstrap meetod

Sageli ei ole selge, millised on õiged eeldused jaotuse kohta, kust andmed pärinevad. Seetõttu on küllalt populaarsed meetodid, mis võimaldavad vähemalt ligikaudselt hinnata huvipakkuvaid suuruseid (hinnangute täpsust, nihet jms) ainult olemasolevate andmete põhjal. Selliseid meetodeid nimetatakse taasvaliku meetoditeks.

12.1 Bootstrap meetod

Sageli on meil antud ainult valim mahuga n ja me ei tea, mis jaotusest see pärineb. Valimi põhjal hindame mingit üldkogumi parameetrit (näiteks dispersiooni, asümmeetriakordajat vms) ning lisaks hinnangule tahame tavaliselt teada hinnangu täpsust (või hinnangufunktsiooni jaotust). Kui me teaks üldkogumi jaotust, siis saaks hinnangu täpsust kindlaks teha näiteks uusi valimeid genereerides. Õiget üldkogumi jaotust me ei tea, kuid me teame valimile vastavat empiirilist jaotust.

Kui juhuslik suurus on jaotusest jatusfunktsiooniga F_X, siis valim on juhuslik vektor jaotusfunktsiooniga

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = F_X(x_1)F_X(x_2)\dots F_X(x_n)$$

Kui kasutame simulatsioonimeetodeid statistiku jaotuse leidmisel, genereerime vektoreid sellest jaotusest. Intuitiivselt, kui muuta jaotusfunktsioone piisavalt vähe, muutub ka enamuse statistikute käitumine küllalt vähe.

Teame, et valimi empiiriline jaotusfunktsioon F_n läheneb valimimahu kasvades jaotusfunktsioonile F_x, seega, asendades F_X funktsiooniga F_n, võib loota, et valimi jaotusfunktsioon ei muutu väga palju. Seetõttu genereerides jaotusfunktsioonile F_n vastavaid sõltumatuid valimeid ja leides huvipakkuva statistiku käitumise, võib loota, et see on lähedane statistiku käitumisele õige jaotuse korral. Jaotusest F_n sõltumatu valimi genereerimine on samaväärne tagasipanekuga juhusliku valimi moodustamisega esialgsesse valimisse sattunud väärtustest.

12.1.1 Nihke hindamine Bootstrap meetodil

Nihke hindamiseks vaatame, kuidas hinnang käitub eeldusel, et valimi empiiriline jaotus vastab üldkogumi jaotusele. Kui me teame üldkogumi jaotust, siis enamasti on võimalik leida huvipakkuva suuruse (parameetri) üldkogumile vastav täpne väärtus, olgu selleks θ_B.

Seejärel leiame statistiku väärtuse m bootstrap valimi korral (tagasipanekuga juhuslik valik esialgse valimi väärtustest, valimi suurus võrdne esialgse valimiga), saame hinnangud

$\hat{\theta}_i,\ i = 1, \ldots, m$

Sageli nii saadud hinnangute keskmine ei koondu empiirilise jaotuse põhjal arvutatud täpseks väärtuseks, st hinnang on nihkega. Nihke hinnanguks võtame $\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i - \theta_B$

Saadud nihke hinnangu abil võime parandada esialgsel valimil statistikut rakendades saadud hinnangut, lahutades selle nihkest maha.

Et saada paremini aru, mis tegelikult tehakse, vaatleme juhtu, kus hindame standardhälvet tavapärase valemiga ja mei valimis on kaks arvu, 1 ja 3. Siis valimi põhjal arvutatud standarhälbe hinanng on $\sqrt{2}$, Bootstrap jaotuse põhjal arvutatud täpne standardhälve on 1 ja kui valida bootstrap jaotusest kaheelemendilisi valimeid ja arvutada nende põhjal saadud standardhälbe hinnangute keskmine, siis piisavalt suure m korral saame (kas oskad põhjendada, miks see nii on?)

$$\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \approx \frac{\sqrt{2}}{2}.$$

Seega bootstrap meetodil leitud nihke suuruseks on $\frac{\sqrt{2}}{2} - 1$ ja parandatud standardhälbe hinnanguks saame $\sqrt{2} + 1 - \frac{\sqrt{2}}{2}$.

12.1.2 Bootstrap usaldusvahemik

Parameetri hinnangu usaldusvahemikke saab bootsrap meetodiga leida mitmel viisil, vaatleme ühte, mis arvestab ka võimalikku nihet. Vaadeldav hinnang põhineb ideel kirjutada otsitav parameeter kujul

$$\theta = \hat{\theta} + (\theta - \hat{\theta}),$$

kus θ on hinnatav parameeter ja $\hat{\theta}$ on valimi põhjal hinnatud väärtus. Leiame vahe $\theta - \hat{\theta}$ jaotuse eeldusel, et lähtejaotus on bootstrap jaotus, st θ asemel kasutame bootstrap jaotuse korral leitud täpset väärtust θ_B. Seega vahe $\frac{\alpha}{2}$-kvantiil on

$$\theta_B - q_{1-\frac{\alpha}{2}},$$

kus q_α tähistab hinnangu $\hat{\theta}$ bootstrap jaotuse korral leitud α-kvantiil:

$$P(\theta_B - \hat{\theta} \leq \theta_B - q_{1-\frac{\alpha}{2}}) = P(\hat{\theta} \geq q_{1-\frac{\alpha}{2}}) = 1 - P(\hat{\theta} < q_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2},$$

Vahe $(1 - \frac{\alpha}{2})$-kvantiil on $\theta_B - q_{\frac{\alpha}{2}}$, seega bootstrap usaldusvahemik olulisuse nivool α on

$$(\hat{\theta} + \theta_B - q_{1-\frac{\alpha}{2}}, \hat{\theta} + \theta_B - q_{\frac{\alpha}{2}}).$$

12.1.3 Parameetriline bootstrap

Lihtsa bootstrap meetodiga ei teki meil kordusvalimeid moodustades kunagi väärtuseid, mida esialgses valimis ei olnud. Kui meil on aga alust arvata, et valim pärineb mingist konkreetsest jaotusest, kuid vaadeldava statsitiku jaoks sobivaid hinnanguid ei ole teoreetiliselt tuletatud, siis on loomulikuks ideeks kasutada simulatsioone sobitatud jaotuse abil, et määrata huvipakkuva parameetri käitumine ja võimalik nihe. Seega protseduur on järgmine:

1. Sobitame valimile jaotuse, st hindame valitud tüüpi jaotuse parameetrid
2. Genereerime kordusvalimeid huvipakkuvast jaotusest ning arvutame nihke, parandatud hinnangu ja usaldusintervalli sarnaselt tavalisele bootstrapile.

Peatükk 13

Loeng 13: Jackknife meetod ja integraalide arvutamine MC meetodil

13.1 Jackknife meetod

Meetodi motiveerimiseks vaatleme suuruse $f(EX)$ hindamist valimi põhjal, kasutades statistikut $S = f(\bar{X})$, kus $\bar{X}$ tähistab valimi aritmeetilist keskmist. Kasutades funktsiooni f Taylori rittaarendust kohal $x = EX$, saame

$$f(\bar{X}) = f(EX) + f'(EX)(\bar{X} - EX) + \frac{f''(EX)}{2}(\bar{X} - EX)^2 + \dots.$$

Seega nihe avaldub kujul

$$E(f(\bar{X}) - f(EX) = \frac{f''(EX)}{2n}DX + \dots$$

ja ei ole üldjuhul võrdne nulliga.

Tähistame nüüd kujul $S(i)$ sama statistikut, mis on hinnatud ilma i-nda vaatluseta ja defineerime uue statistiku valemiga

$$S^*_{(i)} = nS - (n-1)S(i).$$

Rittaarendusi kasutades on nüüd lihtne veenduda, et $S^*_{(i)}$ on väiksema nihkega, st rittaarenduses rohkem liikmeid kaovad ära (veendu selles!). Kuna selliseid hinnanguid saab arvutada valimi iga elemendi ärajätmisel, siis kokkuvõttes peaks kõige parema hinnangu saama siis, kui leida nende keskmine. Seega **Jacknife hinnang** on kujul

$$S^* = \frac{1}{n}\sum_{i=1}^{n} S^*_{(i)} = nS - \frac{n-1}{n}\sum_{i=1}^{n} S(i).$$

Siit saab leida ka jackknife hinnangu nihkele, milleks on

$$nihe_{jack} = S - S^* = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} S(i) - S\right).$$

Lisaks väiksema nihkega hinnangule on enamasti vaja teada ka uue hinnangu usaldusväärsust, st selle standardhälvet ja usaldusintervalle.

Jacknife hinnangu standardhälbe hinnang on

$$s^*_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(S(i) - \frac{1}{n}\sum_{j=1}^{n} S(j)\right)^2} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(S^*_{(i)} - S^*\right)^2}$$

ning Turkey (1958) valem jackknife hinnangu usaldusintervallile olulisusnivool α on kujul

$$(S^* - t_{n-1,1-\frac{\alpha}{2}} s^*_{jack}, S^* + t_{n-1,1-\frac{\alpha}{2}} s^*_{jack}),$$

kus $t_{k,\alpha}$ tähistab vabadusastmete arvuga k t-jaotuse α-kvantiili.

13.2 Integraalide ja kesväärtuste arvutamine MC meetodiga

Sageli pakuvad praktilist huvi teatud juhuslike suuruste keskväärtused, näiteks oodatav tulu investeerimisstrateegialt, finantsoptsioonide hinnad, kindlustusfirma järgmise viie aasta jooksul laostumise tõenäosus jne. Enamasti esituvad huvipakkuvad keskväärtused kujul kujul $g(X)$, kus g on mingi funktsioon ja X on juhuslik suurus või vektor.

Kui X on pidev juhuslik suurus või vektor, vastab keskväärtuse arvutamine (ühe- või mitmekordse) integraali arvutamisele

$$E\, g(X) = \int_{R^m} g(x) f_X(x) dx,$$

kus f_X on juhusliku suuruse (või vektori) X tihedusfunktsioon. Samas integraalidel on palju muid rakendusi, näiteks erinevate objektide pindalad/ruumalad, kehale mõjuva jõu poolt tehtud töö keha liikumisel, muutuva tihedusega keha mass jne.

Tuleb välja, et iga integraali saab esiteda keskväärtusena. Vaatleme integraali

$$I = \int_D f(x) dx,\ D \subset \mathbb{R}^m.$$

Valime ühe tihedusfunktsiooni f_X, mille korral kehtib $f_X(x) > 0,\ x \in D$ ning defineerime

$$g(x) = \begin{cases} \frac{f(x)}{f_X(x)}, & x \in D, \\ 0, & x \notin D. \end{cases}$$

Siis

$$I = \int_D \frac{f(x)}{f_X(x)} f_X(x)\, dx = \int_{R^m} g(x)\, f_X(x) dx = E\, g(X).$$

Jaotuse f_X valikul tuleb aga mõelda sellele, et tekkiv funktsioon g mingite argumentide korral väga halvasti käituma (st mingite x väärtuste korral liiga kiiresti pluss- või miinuslõpmatusse minema) ei hakka, sest halvasti valitud f_X korral võib $g(X)$ dispersioon olla lõpmatu ja siis koondub valimi keskmine keskväärtuseks väga aeglaselt.

13.3 Keskväärtuse hindamine MC meetodil

Vaatleme juhuslikku suuruse $Y = g(X)$ keskväärtuse arvutamist Monte Carlo meetodil. Algoritm on järgmine:

1. Genereerime n-elemendilise valimi X jaotusest, saame $x_1, \ldots, x_n$

2. Rakendame funktsiooni g, saame valimi Y jaotusest: $g(x_1), \ldots, g(x_n)$

3. Hindame Y keskväärtust valimi keskmise abil

$$EY \approx \bar{y} = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

4. **Eeldusel, et Y dispersioon on lõplik**, saame valimi keskmise jaoks erinevad hinnangud (vaja osata ka tuletada tsentraalsest piirteoreemist!)

   (a) ligikaudne usaldusintervall (z_α tähistab standardse normaaljaotuse α-kvantiili)

   $$(\bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{n}}, \bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{n}})$$

   (b) Tõenäosusega $1 - \alpha$ kehtiv (ligikaudne) veahinnang:

   $$-z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{n}}$$

   (c) **Tõenäoline viga** on tõenäosusega $\frac{1}{2}$ kehtiv veahinnang

5. Veahinnangute arvutamisel asendame σ_Y valimi põhjal leitud hinnanguga

13.3.1 MC meetodite võrdlemine

Ühte ja sama integraali või keskväärtust saab MC meetodiga arvutada väga erinevaid lähenemisi kasutades. Näiteks integraalide arvutamisel on võimalik valida lõpmatult palju erinevaid tihedusfunktsioone ja iga valik annab erineva juhusliku suuruse, mille keskväärtust MC meetodil arvutada. Kuigi kõigil juhtudel on sama keskväärtus, võivad dispersioonid olla vägagi erinevad (osadel juhtudel dispersioon võib ka puududa).

Kui me tahame erinevaid meetodeid omavahel võrrelda, siis selleks on mitmeid võimalusi:

- Tööaeg etteantud täpsuse saavutamiseks
- Etteantud täpsuse saavutamiseks vajaminev genereerimiste arv
- täpsus fikseeritud genereerimiste arvu korral

Kuigi tööaeg ettantud täpsuse saavutamisek on kõige olulisem, sõltub see kasutatavast arvutist ja programmikoodi efektiivsusest, mistõttu on seda raske täpselt hinnata. Seetõttu keskendume genereeritavate juhuslike suuruste arvule.

Eeldame järgnevalt, et Y on lõpliku dispersiooniga, siis täpsus etteantud n korral sõltub ainult fikseeritud α väärtusest ja Y standardhälbest. Seega väiksem standardhälve annab suurema täpsuse.

MC meetodi veahinnangu valemist saame, et genereerimiste arv etteantud täpsuse ϵ saavutamiseks on kujul $c_{\alpha,\epsilon} DY$, kus

$$c_{\alpha,\epsilon} = \frac{z^2_{\frac{\alpha}{2}}}{\epsilon^2}.$$

Tööaeg etteantud täpsuse saavutamiseks on seega $t\, c_{\alpha,\epsilon} DY$, kus t on ühe väärtuse genereerimise aeg. Edaspidi keskendume DY suuruse hindamisele erinevate meetodite korral. Oluliselt erinevate meetodite võrdlemisel tuleks võimalusel arvestada ka ühe väärtuse genereerimise aegade erinevust.

Peatükk 14

Loeng 14: Dispersiooni vähendamise meetodid: antiteetilised muutujad ja kontrollmuutujad

14.1 Antiteetilised (*antithetic*) juhuslikud suurused

Antiteetlisteks (vastandlikeks) nimetame juhuslikke suuruseid, mis on sama jaotusega, kuid negatiivselt korreleeritud. Lihtne näide on järgmine: kui $X_1 \sim U(0,1)$, siis $X_2 = 1 - X_1$ puhul on X_1 ja X_2 antiteetilised juhuslikud suurused (veendu selles!).

Üldiselt, kui pideva juhusliku suuruse X_1 tihedusfunktsioon on sümmeetriline μ suhtes (st $f_{X_1}(\mu - a) = f_{X_1}(\mu + a)\ \forall a \in \mathbb{R}$), siis $X_2 = 2\mu - X_1$ korral on X_1 ja X_2 antiteetilised juhuslikud suurused (veendu selles!).

Lisaks on oluline tähele panna, et kui X_1 ja X_2 on sama jaotusega, siis suvalise funktsiooni g korral on $g(X_1)$ ja $g(X_2)$ sama jaotusega, kuid ka antiteetiliste X_1 ja X_2 korral ei pruugi alati olla antiteetilised juhuslikud suurused.

Kui on soovi leida MC meetodiga EY, kus $Y = g(X)$ ja X jaotus on sümmeetriline $x = \mu$ suhtes, siis võime defineerida $X_2 = 2\mu - X$ ja

$$\tilde{Y} = \frac{g(X) + g(X_2)}{2}.$$

Siis $D\tilde{Y} = DY \cdot \frac{1+\rho}{2}$, kus $\rho = \mathrm{cor}(g(X), g(X_2))$. Suuruse $\tilde{Y}$ dispersioon ei ole kunagi suurem, kui DY Kuna aga $\tilde{Y}$ ühe väärtuse arvutamine on umbes 2 korda töömahukam kui Y ühe väärtuse arvutamine, võidame tööajas ainult siis, kui $\rho < 0$ ehk kui $g(X)$ ja $g(X_2)$ on antiteetilised

Eelneva põhjal on meetodi kasulikkuse jaoks väga oluline, et $g(X)$ ja $g(X_2)$ oleks antiteetilised. Osutub, et kui g on monotoonne funktsioon, siis see tingimus on täidetud.

Teoreem 19 *Eeldame, et X pidev juhuslik suurus, mille tihedusfunktsioon on sümmeetriline punkti $x = \mu$ suhtes ning et g on monotoonne (mittekahanev või mittekasvav) funktsioon, mille korral $D(g(X)) < \infty$. Siis*

$$\rho = \mathrm{cor}(g(X), g(X_2)) \leq 0.$$

Tõestus. Teoreemi tõestamiseks piisab näidata, et

$$\mathrm{cov}(g(X), g(X_2)) \leq 0.$$

Paneme tähele, et suvaliste juhuslike suuruste X ja Y ja suvalise reaalarvu a korral kehtibkorral kehtib

$$\begin{aligned}\operatorname{cov}(X,Y) &= E[(X-EX)(Y-EY)] = E[(X-EX)(Y-a)+(X-EX)(a-EY)] \\ &= E[(X-EX)(Y-a)].\end{aligned}$$

Tähistame $\mu_g = E\,g(X)$. Vaatleme mittekahaneva g juhtu. Kuna juhusliku suuruse keskväärtus on tema väärtuste infiimumi ja supreemumi vahel, saame g mittekahanemise tõttu leida leida sellise arvu x^*, et $g(x) \le \mu_g,\ x < x^*$ ja

$$g(x) \ge \mu_g,\ x > x^*.$$

Valime $a = g(2\mu - x^*)$, siis

$$\operatorname{cov}(g(X), g(X_2) = E[(g(X)-\mu_g)(g(X_2)-a)].$$

Kui $X < x^*$, siis x^* definitsiooni kohaselt $g(X)-\mu_g \le 0$. Samas kehtib siis $X_2 = 2\mu - X > 2\mu - x^*$ ning g mittekahanemise tõttu saame

$$g(X_2) \ge g(2\mu - x^*) = a,$$

seega vaadeldaval juhul kehtib

$$(g(X)-\mu_g)(g(X_2)-a) <= 0.$$

Sarnane arutelu annab ka juhul $X > x^*$, et kehtib võrratus $(g(X)-\mu_g)(g(X_2)-a) <= 0$ ning juhul $X = x^*$ on viimane teguritest a definitsiooni kohaselt võrdne nulliga, mistõttu saime, et juhuslik suurus $(g(X)-\mu_g)(g(X_2)-a)$ on alati mittepositiivne. Keskväärtuse monotoonsuse omadusest järeldub nüüd, et vaadeldav kovariatsioon on mittepositiivne ja teoreemi väide kehtib.

Mittekasvava g korral on arutelu analoogiline (mõtle see läbi!). □

14.2 Kontrollmuutujate meetod

Sageli on võimalik koos huvipakkuva suurusega Y genereerida lihtsam juhuslik suurus Z, mille keskväärtus EZ on teada või mille keskväärtust on lihtne arvutada. Seega saame moodustada uue juhusliku suuruse

$$\tilde{Y} = Y - a(Z - EZ),$$

kus a on meie valitud reaalarv. Lihtne on näha, et $E\tilde{Y} = EY$, nii et Y keskväärtuse asemel võime leida $\tilde{Y}$ keskväärtuse. Kordaja a tuleks valida nii, et $\tilde{Y}$ dispersioon on minimaalne, mis taandub ruutvõrrandi (a suhtes) miinimumkoha leidmisele. Lahendades selle võrrandi, saame

$$a = \frac{\operatorname{cov}(Y,Z)}{DZ}.$$

Enamasti hindame a väärtuse valimi põhjal, kasutades kovariatsiooni ja dispersiooni hinnanguid.

Asetades optimaalse a väärtuse $\tilde{Y}$ dispersiooni avaldisse, saame

$$D\tilde{Y} = DY \cdot (1-\rho^2),$$

kus $\rho = \operatorname{cor}(Y,Z)$.

Peatükk 15

Loeng 15: Dispersiooni vähendamise meetodid: olulise valimi meetod ja kihtvalimi meetod

15.1 Olulise valimi meetod

Sageli pakuvad praktikas huvi juhuslikud suurused kujul $Y = g(X)$, kus g omandab suuri (olulisi) väärtuseid piirkonnas kuhu X väärtused satuvad küllalt väikese tõenäosusega. Näiteks finantsmatemaatikas huvitavad investoreid sageli optsioonid, mis toovad omanikule raha sisse ainult juhul, kui aktsiahind oluliselt tõuseb (nt 50% praegusega võrreldes 3 kuu jooksul). Sageli sellisel juhul on dispersioon suur, mis annab idee muuta X jaotust nii, et näeme olulisi väärtuseid rohkem. Osutub, et seda saab alati teha. Nimelt saame valida juhusliku suuruse Z nii, et $f_Z(x) > 0$, kui $g(x)f_X(x) \neq 0$ Kuna alati kehtib võrdus (veendu selles!)

$$E\, g(X) = E[g(Z)\frac{f_X(Z)}{f_Z(Z)}],$$

siis saame valida sellise jaotusega Z, mille korral g olulised väärtused on nähtavad suurema tõenäosusega kui X korral.

Oluline on märkida, et seda ideed saab kasutada ka mitmemõõtmelisel juhul: kui sõltumatute juhuslike suuruste X_1 ja X_2 korral kasutada X_1 asemel suurust Z_1 ja X_2 asemel suurust Z_2, siis

$$E\, g(X1, X_2) = E[g(Z_1, Z_2)\frac{f_{X_1}(Z_1)}{f_{Z_1}(Z_1)}\frac{f_{X_2}(Z_2)}{f_{Z_2}(Z_2)}].$$

Kuna normaaljaotus esineb stohhastilistes mudelites väga sageli, leiame suhte $\frac{f_X(z)}{f_Z(z)}$ juhul $X \sim N(\mu_1, \sigma),\ Z \sim N(\mu_2, \sigma)$. Arvutuste tulemuseks (tee need läbi!) saame

$$\frac{f_X(z)}{f_Z(z)} = e^{\frac{2(\mu_1-\mu_2)z+\mu_2^2-\mu_1^2}{2\sigma^2}}.$$

15.2 Kihtvalimi meetod

Olgu $B_1, \ldots, B_m$ mingi sündmuste täissüsteem (vastastikku välistavad, positiivse tõenäosusega, üks alati toimub igal katsel). Teame, et keskväärtust saab esitada tinglike kesk-

väärtuste kaudu:

$$EY = \sum_{i=1}^{m} P(B_i)E(Y \mid B_i).$$

Seega, kui oskame Y väärtuseid genereerida vastavatest tinglikest jaotustest, võime ühe keskväärtuse asemel arvutada MC meetodiga m keskväärtust

Tähistame $p_i = P(B_i)$ ning olgu Y_{ij} sõltumatud juhuslikud suurused tingliku jaotusega $Y|B_i$. Olgu n kasutatavate juhuslike suuruste arv. Siis tavalise MC meetodi korral on EY hinnangufunktsiooniks

$$H_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

kus Y_i on on sõltumatud suurusega Y sama jaotusega juhuslikud suurused ja seega hinnangufunktsiooni dispersioon on $D(H_n) = \frac{DY}{n}$. Vaatame, kas sama arvu juhuslikke suuruseid kasutades on võimalik saada väiksema dispersiooniga hinnang

Olgu $n_i,\ i = 1, \ldots, m$ sellised positiivsed täisarvud, et $\sum_{i=1}^{m} n_i = n$. Defineerime kihtvalimi hinnangufunktsiooni

$$\tilde{H}_n = \sum_{i=1}^{m} p_i \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}.$$

Leiame kihtvalimi hinnangufunktsiooni dispersiooni

$$D\tilde{H}_n = \sum_{i=1}^{m} \frac{p_i^2}{n_i} D(Y \mid B_i).$$

Põhiline küsimus on, kuidas valida n_i nii, et kihtvalimi hinnangufunktsiooni dispersioon oleks väiksem kui tavalise MC oma?

Teoreem 20 *Olgu $B_i, \ldots, B_m$ sündmsute täissüsteem. Tähistame $p_i = P(B_i)$. Kui $n_i = p_i n$, siis*

$$DH_n - D\tilde{H}_n = \frac{1}{n}\sum_{i=1}^{m} p_i \left(\mu_i - EY\right)^2,$$

kus $\mu_i = E(Y \mid B_i)$.

Eelneva teoreemi kohaselt ei ole seega tõenäosustega proportsionaalse valiku korral kihtvalimi hinnangu dispersioon sama arvu juhuslike suuruste kasutamisel kunagi suurem kui tavalise MC hinnangu dispersioon, st tulemus on enamasti täpsem. Valemist on samuti näha, et mida erinevad on juhusliku suuruse tinglikud keskväärtused erinevad üldisest keskmisest, seda suurem on võit täpsuses, seega kasutatavad sündmused võiksid olla hästi seotud vaadeldava juhusliku suuruse väärtustega.

15.2.1 Kihtvalimi meetodi optimaalsed proportsioonid

Nägime, et valik $n_i = p_i n$ on hea (annab enamasti parema meetodi). Aga kas saab paremini?

Kokku peab kasutatavate juhuslike suuruste summa olema n, seega võime otsida proportsioone q_i, $i = 1, \ldots, m$, mis rahuldavad

$$q_i > 0,\ i = 1, \ldots, m,\ \sum_{i=1}^{m} q_i = 1.$$

Otsime selliseid proportsioone, mille korral valiku $n_i = q_i n$ puhul on $D\tilde{H}_n$ minimaalne. Tähistame

$$\sigma_i^2 = D(Y \mid B_i).$$

Saame tingliku ekstreemumi leidmise ülesande: minimiseerida muutujate $q_1, \ldots, q_m$ suhtes suurust

$$f(q_1, \ldots, q_m) = \frac{1}{n} \sum_{i=1}^{m} \frac{p_i^2 \sigma_i^2}{q_i}$$

tingimustel

$$q_i > 0,\ i = 1, \ldots, m,\ \sum_{i=1}^{m} q_i = 1.$$

Lahendame ülesande Lagrange kordajate meetodil.

Lagrange kordajate meetodil (vt. näiteks Kõrgem matemaatika II loengukonspekti) peavad tingimust $\sum_{i=1}^{m} q_i = 1$ rahuldavates ekstreemumpunktides olema funktsiooni

$$f(q_1, \ldots, q_m) + \lambda \cdot (\sum_{i=1}^{m} q_i - 1)$$

osatuletised muutujate $q_1, \ldots, q_m$ järgi võrduma nulliga ning lisaks peab olema rahuldatud kitsendus $\sum_{i=1}^{m} q_i = 1$. Siit saame võrrandid

$$-\frac{p_i^2 \sigma_i^2}{n q_i^2} + \lambda = 0,\ i = 1, 2, \ldots, m,$$

kust järeldub, et positiivsetele proportsioonidele vastavad ekstreemumpunktid esituvad kujul

$$q_i = \frac{p_i \sigma_i}{\sqrt{n\lambda}}$$

Tingimusest $\sum_{i=1}^{m} q_i = 1$ järeldub nüüd, et

$$\sqrt{n\lambda} = \sum_{i=1}^{m} p_i \sigma_i,$$

mistõttu

$$q_i = k p_i \sigma_i,\ i = 1, \ldots, m,$$

kus $k = \frac{1}{\sum_{i=1}^{m} p_i \sigma_i}$.

Seega parima tulemuse saame, kui igas kihis on kasutatud väärtuste arv proportsionaalne korrutisega $p_i \sigma_i$. Hinnangu dispersioon on sel juhul

$$D\tilde{H}_n = \frac{\left(\sum_{i=1}^{m} p_i \sigma_i\right)^2}{n}$$

Kuna iga kihtvalimi meetodi puhul on Suure n korral on hinnang ligikaudu normaaljaotusega (iga kihi keskmine on ligikaudu normaaljaotusega tsentraalse piirteoreemi kohaselt ning sõltumatute normaaljaotusega juhuslike suuruste summa on normaaljaotusega), seega saame ligikaudu leida tõenäosusega $1 - \alpha$ kehtiva vea kujul

$$-z_{\frac{\alpha}{2}} \sqrt{D\tilde{H}_n}$$

15.2.2 Kihtvalimi meetodi rakendamine

Kõigepealt tuleb valida m ja fikseerida sündmused $B_1, \ldots, B_m$. Teoreetilisetl, mida suurem m, seda suuremat võitu on võimalik saada. Praktiliset peame iga kihi korral hindama tinglikke dispersioone ja normaalse täpsusega hinnangu jaoks on vaja piisavalt suur arv (nt 100) genereeritud väärtust, seetõttu minimaalne n_i võiks olla vähemalt 100. Suure m korral aga võib see tähendada väga suurt n väärtust.

Sündmused B_i peaks olema juhusliku suuruse käitumisega seotud. Kui $Y = g(X)$, siis on hea valik $B_i = \{a_{i-1} < X \leq a_i\}$, kus

$$-\infty = a_0 < a_1 < \cdots < a_m = \infty,$$

sest selliste juhuslike suuruste korral oskame me jaotusfunktsiooni pöördfunktsiooni meetodil juhuslike suuruste väärtuseid tinglikust jaotusest genereerida. Sageli on mõistlik defineerida a_i juhusliku suuruse X kvantiilide kaudu, siis saame määrata lihtsalt sündmuste tõenäosused (nt tagada, et $p_i = \frac{1}{m}\ \forall i$).

Kasutatavad tulemused tõenäosusteooriast

Teoreem 21 *(Tõenäaosuse omadused) Olgu* $(\Omega, \mathcal{F}, P)$ *mingi tõenäosusruum. Siis kehtivad järgnevad omadused:*

1. $P(\emptyset) = 0$;
2. *kui* $A_i \in \mathcal{F},\ i = 1, 2, \ldots, n$ *on vastastikku välistavad, st* $A_i \cap A_j = \emptyset,\ i \neq j$, *siis kehtib võrdus*
$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i);$$
3. *kui* $A, B \in \mathcal{F},\ A \subset B$, *siis* $P(A) \leq P(B)$ *(monotoonsus).*
4. $P(\bar{A}) = 1 - P(A)$;
5. $P(A \setminus B) = P(A) - P(A \cap B)\ \forall A, B \in \mathcal{F}$;
6. $P(A) \leq 1\ \forall A \in \mathcal{F}$;
7. $P(A \cup B) = P(A) + P(B) - P(A \cap B)\ \forall A, B \in \mathcal{F}$;
$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n),\ A_i \in \mathcal{F},\ i = 1, 2, \ldots, n$;
8. $P(A \cup B) \leq P(A) + P(B)\ \forall A, B \in \mathcal{F}$;
$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i),\ A_i \in \mathcal{F},\ i \in \mathbb{N}$;
9. *Tõenäosuse pidevus:*
   - $A_i \in \mathcal{F},\ i \in \mathbb{N},\ A_1 \subset A_2 \subset A_3 \subset \ldots \Rightarrow \lim_{n\to\infty} P(A_n) = P(\bigcup_{i=1}^{\infty} A_i)$;
   - $A_i \in \mathcal{F},\ i \in \mathbb{N},\ A_1 \supset A_2 \supset A_3 \supset \ldots \Rightarrow \lim_{n\to\infty} P(A_n) = P(\bigcap_{i=1}^{\infty} A_i)$;

Lemma 22 (Täistõenäosuse valem). *Rahuldagu sündmused* $B_i \in \mathcal{F},\ i = 1, \ldots, n$ *tingimusi*

$$P(B_i) \neq 0,\ i = 1, 2, \ldots, n; \quad B_i \cap B_j = \emptyset, \quad i \neq j; \quad \bigcup_{i=1}^{n} B_i = \Omega. \tag{15.1}$$

Siis iga sündmuse $A \in \mathcal{F}$ *korral kehtib võrdus*

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A|B_i).$$

Definitsioon 23 *Juhusliku suuruse X jaotusfunktsiooniks nimetatakse funktsiooni*

$$F(x) = P(\{X \leq x\}), \quad x \in \mathbb{R}.$$

Teoreem 24 *Olgu X juhuslik suurus ning F tema jaotusfunktsioon. Siis kehtivad järgnevad omadused.*

1. $0 \leq F(x) \leq 1$ *iga* $x \in \mathbb{R}$ *korral.*
2. F *on monotoonselt kasvav: kui* $x_1 < x_2$, *siis* $F(x_1) \leq F(x_2)$.
3. *Kehtivad piirväärtused*
$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$$
4. F *on paremalt pidev:*
$$\lim_{x > a, x \to a} F(x) = F(a) \quad \forall a \in \mathbb{R}.$$
5. *Kehtib võrdus*
$$P(\{X = a\}) = F(a) - \lim_{x < a, x \to a} F(x).$$
6. *Kehtib võrdus* $P(\{a < X \leq b\}) = F(b) - F(a)$.

Definitsioon 25 *Juhusliku vektori (X, Y) jaotusfunktsiooniks (ehk juhuslike suuruste X ja Y ühisjaotuse jaotusfunktsiooniks) nimetatakse funktsiooni*

$$F_{X,Y}(x, y) = P(\{X \leq x,\, Y \leq y\}), \quad x, y \in \mathbb{R}.$$

Definitsioon 26 *Juhuslikku vektorit (X, Y) nimetatakse* pidevaks, *kui tema jaotusfunktsioon avaldub kujul*

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{X,Y}(u, v)\, dv \right) du, \quad x, y \in \mathbb{R}$$

mingi funktsiooni $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ korral. Funktsiooni $f_{X,Y}$ nimetatakse sel juhul juhusliku vektori (X, Y) tihedusfunktsiooniks *(ehk juhuslike suuruste X ja Y* ühistiheduseks*).*

Lemma 27 *(Tihedusfunktsiooni omadused) Olgu (X, Y) pidev juhuslik vektor jaotusfunktsiooniga $F_{X,Y}$ ja tihedusfunktsiooniga $f_{X,Y}$. Siis kehtivad järgmised omadused:*

1. *Funktsioon $f_{X,Y}$ on mittenegatiivne, st $f_{X,Y}(x, y) \geq 0$ $\forall (x, y) \in \mathbb{R}^2$;*
2. *kehtivad võrdused*
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx,$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

3. *Kui* $D \subset \mathbb{R}^2$ *on Boreli* σ*-algebra suhtes mõõtuv hulk (st esitatav loenduva arvu ristkülikute abil kasutades ühendeid, ühisosasid ja täiendeid), siis*

$$P(\{(X,Y) \in D\}) = \iint\limits_D f_{X,Y}(x,y)\,dx\,dy.$$

4. *Kui* $g : \mathbb{R}^2 \to \mathbb{R}$ *on "piisavalt heade omadustega" funktsioon (nt pidev või selline, mille valemit me oskame kirja panna) ning*

$$\iint\limits_{\mathbb{R}^2} |g(x,y)| f_{X,Y}(x,y)\,dx\,dy < \infty,$$

siis

$$E(g(X,Y)) = \iint\limits_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\,dx\,dy.$$

5. *Kui* $F_{X,Y}$ *on diferentseeruv punktis* (x,y)*, siis*

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$$

Lemma 28 *Pidevad juhuslikud suurused* X *ja* Y *on sõltumatud parajasti siis, kui nende ühisjaotuse tihedusfunktsioon avaldub kujul*

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)\ \forall x, y \in \mathbb{R}.$$

Kirjandus

[1] Hull, T.E; Dobell, A. R., *Random Number Generators*, SIAM Review **4** (3), 1962, 230-254

[2] T. Kollo. Monte Carlo Meetodid. Tartu : Tartu Ülikooli Kirjastus, 2004.

[3] Marsaglia, George. "Random Numbers Fall Mainly in the Planes". PNAS. 61 (1): 25–28, 1968