TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

LASSE ENÄSUO

MONTE CARLO -MENETELMIEN HYÖDYNTÄMINEN AINEISTON TUTKIMISESSA

Kandidaatintyö

Tarkastaja: Yliopisto-opettaja Jussi Kangas
Matematiikan laboratorio
15.9.2017

TIIVISTELMÄ

LASSE ENÄSUO: Monte Carlo -menetelmien hyödyntäminen
aineiston tutkimisessa
Tampereen teknillinen yliopisto
Kandidaatintyö, 30 sivua, 1 liite
15.9.2017
Teknisluonnontieteellinen koulutusohjelma
Pääaine: Matematiikka
Tarkastaja: Yliopisto-opettaja Jussi Kangas
Avainsanat: Monte Carlo -menetelmä, Permutaatiotesti, P-arvo, normaalijakauma

Tässä työssä tutkitaan Monte Carlo -menetelmien ominaisuuksia ja hyödyntämismahdollisuuksia. Hyvin usein aineistojen tutkiminen vaatii, että aineisto noudattaa jotain todennäköisyysjakaumaa. Monte Carlo -menetelmiä käytettäessä aineiston ei kuitenkaan tarvitse noudattaa tunnettuja todennäköisyysjakaumia, vaan aineistoja pystytään tutkimaan muodostamalla niistä yksinkertaisia mallinnusohjelmia, jotka toimivat toistokokeen omaisesti. Työssä tutustutaan pintapuolisesti muutamaan menetelmään, ja sen lisäksi Monte Carlo -permutaatiotestiä tutkitaan tarkemmin. Permutaatiotestistä luodaan tilastollinen testausohjelma, jolla voidaan tutkia erityyppisiä aineistoja.

Permutaatiotestin perusidea on vertailla itse valittua vertailuryhmää koko muun populaation. Testissä arvotaan aina vertailuryhmän alkioiden lukumäärän verran alkioita koko otospopulaatiosta ja verrataan näistä laskettua keskiarvoa vertailuryhmän keskiarvoon. Lopuksi testin antaman P-arvon perusteella voidaan päätellä, onko vertailuryhmän keskiarvo satunnaisuus huomioiden todella suurempi kuin koko otospopulaation keskiarvo.

Permutaatiotestin mallinnusohjelma osoittautuu toimivaksi tavaksi tutkia aineistoja. Testi antaa vain vähän toisistaan poikkeavia P-arvoja, kun toistomäärä yksittäisessä testissä pidetään riittävän suurena. Permutaatiotesti on lisäksi tietokoneelle hyvin kevyt suorittaa, joten testi voidaan helposti toistaa monta kertaa. Siten käyttäjä pystyy itse tutkimaan P-arvon käyttäytymistä ja tekemään siitä laajempia johtopäätöksiä. Permutaatiotesti osoittautuu siis hyödylliseksi välineeksi erilaisten aineistojen tutkimisessa riippumatta siitä, että noudattaako aineisto mitään tunnettua todennäköisyysjakaumaa vai ei.

SISÄLTÖ

LYHENTEET JA MERKINNÄT

A	Permutaatiotestin vertailuryhmän joukko
B	Permutaatiotestin otospopulaation joukko
MC	Monte Carlo
$n(A)$	Vertailuryhmän alkioiden lukumäärä
$n(B)$	Otospopulaation alkioiden lukumäärä

α	Riskitaso tilastollisessa testauksessa
μ	Odotusarvo
σ	Keskihajonta
σ^2	Varianssi
$\overline{x}$	Keskiarvo

1. JOHDANTO

Tässä työssä tutkitaan erilaisten Monte Carlo -menetelmien ominaisuuksia sekä eri menetelmien hyödyntämisen mahdollisuutta käytännön elämässä. Yleinen ongelma erilaisten aineistojen tarkastelussa syntyy silloin, kun aineisto ei noudata mitään tunnettua todennäköisyysjakaumaa. Tämän seurauksena ongelman mallintaminen on haastavaa eikä luotettavaa arviota populaation käyttäytymisestä pystytä antamaan.

Monte Carlo -menetelmät, erityisesti permutaatiotesti, soveltuvat hyvin tämänkaltaisten ongelmien ratkaisemiseen, koska nämä menetelmät ovat perusidealtansa hyvin yksinkertaisia. Tämän seurauksena ne ovat myös usein kevyitä suorittaa tietokoneohjelmilla. Kun monimutkaiseen ongelmaan löydetään sopiva ja tehokas menetelmä, jossa voidaan hyödyntää tehokkaasti tietokoneen laskentatehoa, saadaan yleensä hyvin lyhyessä ajassa ratkaistua ongelma numeerisesti.

Tässä työssä perehdyään erityisesti Monte Carlo -permutaatiotestiin ja sen ominaisuuksiin. Permutaatiotestistä luodaan lisäksi tilastollinen testausohjelma, johon käyttäjä pystyy syöttämään omaa dataa tarkasteltavaksi. Käyttäjä valitsee käsiteltävästä aineistosta vertailuryhmän, jonka alkioiden arvoja mallinnusohjelma vertaa koko populaation alkioiden arvoihin. Toistamalla tätä riittävän monta kertaa saadaan mahdollinen vertailuryhmän ja koko populaation välinen eroavaisuus selville. Testausohjelman avulla käyttäjä pystyy siis tutkimaan ja tulkitsemaan sellaista aineistoa, joka ei ole laskettavissa helposti eikä noudata tunnettuja todennäköisyysjakaumia.

2. MONTE CARLO -MENETELMIEN ESITTELY

Tässä luvussa tutustutaan erilaisiin Monte Carlo -menetelmiin, menetelmien käyttöön liittyviin vaatimuksiin sekä niiden taustalla olevaan teoriaan. Monte Carlo -menetelmät (MC -menetelmät) sopivat erityisen hyvin sellaisiin tilanteisiin, joissa tutkittavan aineiston todennäköisyysjakauma ei noudata mitään tunnettua todennäköisyysjakaumaa tai ongelma on hyvin monimutkainen ratkaista [1, s. 336]. MC -menetelmiä käytetään paljon fysiikan- ja kemianalan ilmiöiden mallintamiseen sekä matemaattisten ongelmien ratkaisemiseen [2, s. 7–19]. MC -menetelmiä käytettäessä pitää kuitenkin muistaa, että saadut tulokset ovat aina numeerisia ratkaisuja ja virhetarkkuus riippuu monista asioista. Virhetarkkuuteen vaikuttavat esimerkiksi käytettävän menetelmän suoritustapa, menetelmän vaativuus, otoskokoko ja tarkasteltavien arvojen suuruus.

2.1 Taustatietoa Monte Carlo -menetelmistä

Kaikki Monte Carlo -menetelmät perustuvat samaan perusideaan. MC -menetelmissä tiettyä kaavaa toistetaan lukuisia kertoja, jolloin saadaan numeerinen arvio tutkittavalle asialle [3, s. 1–8]. Tästä seuraa myös MC -menetelmiä kuvailtaessa käytetty englanninkielinen termi ”resampling”. Ratkaisutavasta riippuen tietokoneen laskentatehokkuus riittää ratkaisemaan ongelman hyvin lyhyessä ajassa. Tietokoneiden laskentatehokkuuden kasvaessa yhä suurempia ongelmia pystytään ratkaisemaan järkevässä ajassa. Täten myös toistomääriä yhdessä testissä pystytään nostamaan laskentatehokkuuden kasvaessa, mikä mahdollistaa entistä tarkemmat numeeriset tulokset. Menetelmien mallintamisessa tärkeää on siis tehdä toimivan mallin lisäksi itse mallista tehokas, sillä liian raskas ohjelma kuluttaa paljon laskenta-aikaa. Tulevaisuudessa seuraava suuri askel eteenpäin on kvanttitietokoneiden kehittäminen toimivaksi kokonaisuudeksi, jolloin laskentatehoa saadaan lisää yhä vaikeampien ongelmien ratkaisemiseen.

Monte Carlo -menetelmissä on yksi toimintaperiaatteeseen liittyvä ongelma. Ongelma liittyy satunnaismuuttujien generoimiseen. Monet MC -menetelmistä ratkais-

taan tietokoneen avulla, joten ongelmaksi muodostuu se, että kuinka pystytään tietokoneohjelman avulla luomaan täysin satunnaisia tapahtumia. Tietokoneohjelmat arpovat satunnaisluvut jollakin algoritmilla, joten miten voidaan varmistaa se, että tapahtumat ovat täysin satunnaisia. Jun ja Jöckel [1][2] kertovat tämän olevan todellinen ongelma, sillä virheellinen satunnaislukujen arpominen voi vaikuttaa merkittävästi lopputulokseen. Ongelman laajuuden ja sen monimutkaisuuden vuoksi tässä työssä ei tarkastella kyseistä ongelmaa, vaan työssä oletetaan, että mallinnusohjelma osaa arpoa täysin satunnaisesti mallin vaativat asiat. Myös virhetarkastelussa jätetään tämä asia huomioimatta.

Monte Carlo -menetelmiä käytetään monien fysikaalisten ja kemiallisten ilmiöiden mallintamiseen. MC-menetelmät ovat hyviä tapoja mallintaa sellaisia tilanteita, joissa tarkasteltava ilmiö tapahtuu hyvin pienessä koossa, esimerkiksi atomitasolla. Näissä tapauksissa ilmiön mallintaminen on hankalaa, joten MC-menetelmien numeeriset ratkaisut ilmiön mallintamiseksi ovat yleensä riittävän tarkkoja, kunhan mallinnusmenetelmä on toimiva. Hyvä esimerkki MC-menetelmien hyödyntämisestä tutkimuksessa on se, että Leino kykenee omassa diplomityössään [4] mallintamaan jopa kvanttipisteitä käyttäen hyväksi Monte Carlo -polkuintegraaleja. MC -menetelmien hyödyntäminen ei siis rajoitu ainoastaan matemaattisten ongelmien ratkaisemiseen. MC -menetelmiä käytetään paljon fysikaalisten ilmiöiden taustalla esiintyvässä laskennassa, molekyylien rakenteiden mallintamisessa sekä aineistojen tutkimisessa [2, s. 7–22][4]. MC -menetelmistä saatava hyöty perustuukin siihen, että erityyppisiin ongelmiin voidaan hyödyntää hyvinkin erilaisia Monte Carlo -menetelmiä.

2.2 Menetelmissä käytettävä matematiikka

MC -menetelmissä käytetään paljon todennäköisyysmatematiikkaa hyödyksi. Keskeisiä ydinkäsitteitä useille menetelmille ovat todennäköisyyslaskennasta tutut käsitteet tiheysfunktio, keskiarvo $\overline{x}$, varianssi σ^2, keskihajonta σ ja odotusarvo μ. Näitä käsitteitä hyödynnetään MC -menetelmien haastavampien kaavojen johtamisessa tai lopputulosten tulkinnassa. Tässä työssä pääpaino on Monte Carlo -permutaatiotestin mallintaminen sekä mallin tulkintaan ja toimintaan liittyvät asiat, joten tässä työssä muita MC -menetelmiä ei käydä niin yksityiskohtaisesti lävitse.

Monissa MC -menetelmissä hyödynnetään keskiarvoa, joten määritetään keskiarvo $\overline{x}$ seuraavasti:

Määritelmä 1 . *Olkoon $[x_1, x_2, \ldots, x_k]$ tarkasteltavan joukon S alkioita ja olkoon n joukon S alkioiden lukumäärä. Tällöin keskiarvolle $\overline{x}$ saadaan lauseke*

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

Siirtämällä yhteinen tekijä $\frac{1}{n}$ murtoluvun vasemmalle puolelle saadaan keskiarvolle muodostettua seuraava summakaava.

Lause 1 *Joukolle $S = [x_1, x_2, \ldots, x_k]$ voidaan määrittää keskiarvo $\overline{x}$ seuraavalla summakaavalla*

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

missä n kuvaa joukon S alkioiden lukumäärää.

Usean MC -menetelmän toiminta perustuu siihen, että verrataan tuntematonta asiaa tunnettuun asiaan. Jos tunnetun vertailukohteen odotusarvo μ ja varianssi σ^2 tunnetaan, niin näitä voidaan hyödyntää usein myös tuntemattoman asian tutkimisessa. Tämän seurauksena odotusarvo ja varianssi ovat tärkeitä työkaluja monissa MC -menetelmissä. Odotusarvo voidaan määritellä seuraavasti [3, s. 9].

Määritelmä 2 *Olkoot X diskreetti satunnaismuuttuja ja S_X satunnaismuuttujan otosavaruus. S_X sisältää alkiot $[x_1, x_2, \ldots, x_k]$. Olkoon p_i se todennäköisyys, että satunnaismuuttuja X saa arvon x_i. Tällöin satunnaismuuttujalle saadaan odotusarvo*

$$E(X) = \sum p_i x_i = \mu$$

Usein pelkkä odotusarvo ei riitä MC -menetelmissä laajempien johtopäätösten tekemiseen, vaan lisäksi tarvitaan tietoa siitä, että kuinka paljon käsiteltävän aineiston arvot eroavat toisistaan. Aineiston arvojen vaihtelevuutta voidaan kuvata hyvin varianssin σ^2 avulla. Mitä suurempi varianssi on, niin sitä kaempana odotusarvosta arvot ovat keskimäärin. Varianssi määritetään odotusarvoa hyödyntäen seuraavasti [5, s. 35].

Määritelmä 3 *Olkoot X satunnaismuuttuja, μ sen odotusarvo ja Y=$(X - \mu)^2$. Tällöin varianssi on funktion Y odotusarvo $E(Y)$ eli*

$$Var(X) = E((X - \mu)^2)$$

Varianssi voidaan määritellä myös lauseessa 2 esitetyllä tavalla [3, s. 11], jota pystytään yleensä paremmin hyödyntämään kuin määritelmän 3 mukaista kaavaa.

Lause 2 *Olkoot X satunnaismuuttuja ja $E(X)$ satunnaismuuttujan odotusarvo. Tällöin varianssi voidaan ilmoittaa muodossa*

$$Var(X) = E(X^2) - E(X)^2 = \sigma^2$$

Todistus. Olkoot X satunnaismuuttuja ja $E(X)$ satunnaismuuttujan odotusarvo. Merkitään $E(X) = \mu$. Määritelmän 3 mukaan varianssi on

$$Var(X) = E((X - \mu)^2)$$

Aukaistaan ensiksi sulut.

$$Var(X) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2$$

Käytetään hyödyksi merkintätapaa $E(X) = \mu$ yhtälön sieventämisessä. Yhtälö saadaan lopulta muotoon

$$E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - E(X)^2 \qquad \square$$

Lisäksi tietyissä tapauksissa MC -menetelmien tuloksien tulkinnassa saatetaan tarvita myös keskihajontaa, joka saadaan ratkaistua helposti ottamalla varianssista neliöjuuri [5, s. 29].

Määritelmä 4 *Olkoot X satunnaismuuttuja ja $E(X)$ satunnaismuuttujan odotusarvo. Nyt keskihajonta $D(X)$ saadaan muotoon*

$$D(X) = \sqrt{Var(X)} = \sqrt{\sigma^2} = \sigma$$

2.3 Monte Carlo integraalit ja estimaattorit

Monte Carlo -menetelmien avulla voidaan laskea monia integraalien arvoja hyvin tarkasti. Näissä menetelmässä hyödynnetään yleensä jo määritelmässä 2 olevaa odotusarvoa $E(X) = \sum p_i x_i$. Yksi mahdollinen tapa laskea integraalin arvo MC -integroimismenetelmällä on pilkkoa integraalialue pieniin osiin. Tämän jälkeen lasketaan erikseen jokaisen pienen osan integraali esimerkiksi tietokoneohjelmalla, ja lopuksi summataan lasketut arvot yhteen, jolloin saadaan selville alkuperäisen

integraalin arvo. MC-integroimismenetelmää käyttämällä integrointi on hyvin nopeaa ja laskettavien termien määrä on suurin laskentanopeuteen vaikuttava tekijä [3, s. 28]. Tämä menetelmä antaa poikkeuksellisesti tarkan integraalin arvon, koska integroimissääntöjen mukaan integraali voidaan jakaa osiin seuraavasti

$$\int_a^d f(x)\,dx = \int_a^b f(x)\,dx + \ldots + \int_c^d f(x)\,dx, \qquad a \leq b \leq \ldots \leq c \leq d$$

jos integroitava funktio $f(x)$

$$\int_a^d f(x)\,dx \tag{2.1}$$

on jatkuva integroimisrajojen sisällä. [6, s. 115]

Edellä mainitulla menetelmällä saadaan ratkaistua tarkkoja arvoja, mutta menetelmää ei käytetä ongelman ratkaisemisessa, koska se ei välttämättä ole käyttökelpoinen tietyissä tilanteissa tai likiarvo on riittävä ongelman ratkaisemiseksi. MC -menetelmien avulla voidaan laskea esimerkiksi satunnaislukuja antavasta funktiosta g odotusarvo μ, jonka avulla päästään usein käsiksi funktion varianssiin σ^2 ja sitä kautta funktion muihin ominaisuuksiin. Funktion g antamat satunnaisluvut voivat olla painotettuja, ja siitä huolimatta MC -menetelmillä pystytään ratkaisemaan odotusarvo, mikäli funktion g tiheysfunktio tiedetään. Jotta funktio f on satunnaismuuttujan X tiheysfunktio, niin sen pitää toteuttaa määritelmässä 5 olevat ehdot [5, s. 25][6, s. 121]. Määritellään tiheysfunktio seuraavasti:

Määritelmä 5 . *Funktio $f : \mathbb{R} \to \mathbb{R}$ on satunnaismuuttujan X tiheysfunktio, jos seuraavat ehdot toteutuvat.*

1. $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)dx, \quad a \leq b$

Jos tiedetään funktion g tiheysfunktio, saadaan funktion g odotusarvolle G arvio

käyttämällä hyödyksi lauseen 1 keskiarvon summakaavaa

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

Funktion $g(x)$ satunnaismuuttujan arvot saattavat olla painotettuja, joten tästä syystä odotusarvoa ei voida suoraan laskea määritelmän 2 perusteella. Oletetaan, että funktio $g(x)$ antaa diskreettejä satunnaismuuttujia. Siten $g(x)$:n tiheysfunktio $f(x)$ kertoo satunnaismuuttujan eri arvojen esiintymistodennäköisyyden. Kun siirretään keskiarvon kaavassa satunnaisluvun todennäköisyyttä kuvaava termi $\frac{1}{n}$ summakaavan sisälle, ja korvataan se funktion $g(x)$ tiheysfunktiolla $f(x)$, saadaan funktion g odotusarvolle G arvio integroimalla tiheysfunktion $f(x)$ ja funktion $g(x)$ tuloa tiheysfunktion määrittelyjoukon ylitse. Tällöin odotusarvolle saadaan yhtälö

$$G = \int_{-\infty}^{\infty} f(x)g(x)\, dx \tag{2.2}$$

Todistus sivuutetaan, koska se on haastava ja käsiteltävä aihe sivuaa työn pääaihetta eli permutaatiotestiä. Integraalin arvo pystytään ratkaisemaan numeerisesti MC -integroimismenetelmällä käyttäen hyväksi satunnaislukuja. Tätä menetelmää voidaan hyödyntää, kun halutaan selvittää funktion g integraali itse valittujen integroimisrajojen sisällä. Tässä menetelmässä arvotaan satunnaisluku integroimisrajojen välistä ja lasketaan funktion g arvo kyseisessä pisteessä [3, s. 28–32]. Tätä menettelytapaa toistetaan N kertaa ja summataan saadut g:n arvot, minkä jälkeen saatu summa kerrotaan integroimisrajojen erotuksen ja toistojen lukumäärän osamäärällä. Kun tarkastellaan integraalia arvoa välillä $[a, b]$, voidaan MC -menetelmällä saada integraalin likiarvo I selville seuraavaa kaavaa hyödyntäen [7, s. 5].

$$I = \frac{b-a}{N} \sum_{i=1}^{N} f(X_i), \tag{2.3}$$

missä N kertoo toistojen lukumäärän ja X on satunnaismuuttuja, joka arvotaan integroimisrajojen sisältä. Selvitetään MC -integroimisessa tapahtuvan virheen suuruus tutkimalla tapausta, jossa $N \to \infty$. Tällöin virheen suuruus lähestyy nollaa ja siten pätee:

$$G = \frac{1}{N} \sum_{i=1}^{N} g(x_i) = \int_{-\infty}^{\infty} f(x)g(x)\, dx\ , \quad kun\ N \to \infty$$

Yhtälön vasemmasta puolesta saadaan varianssi selville toistojen lukumäärän N suh-

teen, kun sijoitetaan lauseesta 2 saatu varianssi eli $Var(g(x_i)) = \sigma^2$ yhtälön vasemmalla puolella olevan $g(x_i)$:n paikalle. Täten saadaan

$$Var(\frac{1}{N}\sum_{i=1}^{N} g(x_i)) = \frac{\sigma^2}{N} \tag{2.4}$$

Tästä voidaan päätellä, että otoskoon N kasvaessa integraalin arvon numeerisen ratkaisun varianssi pienenee. Numeerisen integraalin arvon varianssi pienenee siis suhteessa $\frac{1}{N}$. Koska määritelmän 4 mukaan keskihajonta $D(X) = \sqrt{\sigma^2}$, niin saadun integraalin likiarvon keskihajonta pienenee suhteessa $\frac{1}{\sqrt{N}}$.

Integraalien saamat arvot voivat vaihdella merkittävästi riippuen siitä, että mihin kohtaan integroimisväliä satunnaisluku osuu. Tämän takia kyseinen MC -integroimismenetelmä vaatii erittäin korkeat toistomäärät, jotta saatu likiarvo on mahdollisimman tarkka. Tämän seurauksena tulee N:ksi valita mahdollisimman suuri luku siten, että tietokoneohjelman laskenta-ajat pysyvät vielä järkevinä. Koska helppojen ja yksinkertaisten integraalien laskeminen on mahdollista myös lukuisilla muilla tavoilla, niin tätä satunnaislukuja hyödyntävää Monte Carlo -integroimistapaa kannattaa käyttää vaikeiden integraalien tapauksissa. Vaikeita integraaleja voivat esimerkiksi olla pintojen ja tasojen väliset integraalit, joissa on yleensä useampi muuttujaa, ja siten integraalit ovat monimutkaisempia ja haasteellisimpia ratkaista muilla menetelmillä.

Toisaalta MC -integroinnin etuna on sen yksinkertaisuus. Tämä menetelmä on yleensä tietokoneelle kevyt suorittaa menetelmän yksinkertaisuuden takia, joten toistojen lukumäärä N saadaan yleensä suureksi ilman, että ohjelman suorittaminen kuluttaa tarpeettomasti aikaa. Näin vaikeidenkin integraalien tapauksessa tämä MC -integroimismenetelmä antaa yleensä riittävän tarkkoja lopputuloksia.

Integraalien likiarvoja voidaan laskeja kuvaajista MC -menetelmillä arpomalla kuvaajista koordinaattipisteitä [3, s. 29–32]. Tällä menetelmällä voidaan laskea muun muassa pinta-aloja tai tilavuuksia kuvaajista, jos koordinaattiakseleiden skaalaukset ovat merkitty kuvaan. Tämä menetelmä perustuu siihen, että arvotaan satunnaisluvut koordinaattiakselien minimi- ja maksimiarvojen välistä, ja tutkitaan sitä, että osuuko arvotut pisteet kuvaajan tarkasteltavaan pinta-alaan tai tilavuuteen. Menetelmä on hyvin yksinkertainen, sillä koordinaattiakselien minimi- ja maksimiarvoista voidaan laskea koko kuvaajan pinta-ala tai tilavuus. x, y -koordinaatistossa olevan kuvaajan peittämäksi pinta-alaksi saadaan siis:

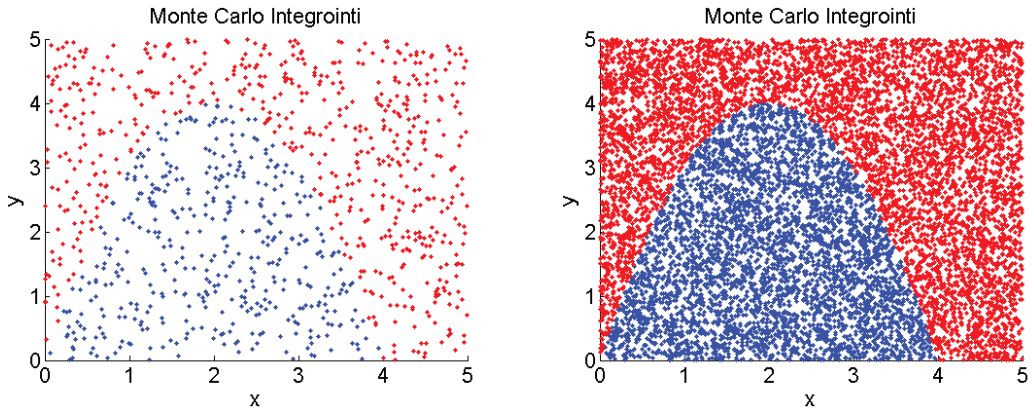
$$A_{kuvaaja} = |\Delta(x_{max} - x_{min}) \cdot \Delta(y_{max} - y_{min})| \tag{2.5}$$

Kun menetelmää on toistettu haluttu määrä kertoja, niin lasketaan pinta-ala arvottujen koordinaattipisteiden osumissuhteen avulla eli integraalin arvoksi saadaan:

$$A_{integraali} = \frac{Osumat}{N} \cdot A_{kuvaaja} \ , \tag{2.6}$$

jossa N on arvottavien koordinaattipisteiden lukumäärä ja $A_{integraali}$ on halutun integraalin likimääräinen pinta-ala.

Tämän menetelmän toimintatapaa voidaan havainnollistaa kuvan 2.1 avulla, jossa funktion $f(x) = -x^2 + 4x$ integraalin arvosta välillä $[0, 4]$ saadaan likiarvo käyttämällä hyödyksi satunnaislukujen arvonta -menetelmää.



Kuva 2.1 *"Monte Carlo integroiminen satunnaislukujen arvonta -menetelmällä*

Vasemmanpuoleisessa kuvassa arvontapisteitä on 1000 ja oikeanpuoleisessa kuvassa niitä on 10000. Kuvasta 2.1 huomataan, että 1000:lla pisteellä saadaan jo kohtuullisen tarkka tulos, sillä arvotut pisteet peittävät melko tasaisesti koko kuvaajan. Arvontapisteiden lukumäärän ollessa 10000 saadaan erittäin tarkasti peitetty kuvaaja, joten integraalin likiarvon tarkkuus parantuu huomattavasti arvontapisteitä kasvattamalla. Toki melko tarkan likiarvon saa jo 1000:lla pisteella, mutta käyttämällä riittävää määrää arvontapisteitä päästään jo hyvin lähelle interaalin oikeaa arvoa $\frac{32}{3} \approx 10{,}67$. Mitä enemmän arvontapisteitä käytetään, niin sitä vähemmän integraalin arvo vaihtelee testiä toistettaessa.

Taulukko 2.1 *Monte Carlo -integroiminen kuvaajasta*

arvontapisteitä	$\int_{min}$	$\int_{max}$	keskiarvo	keskiarvon virhe
100	9,250	12,250	10,950	0,283
1000	10,200	11,275	10,743	0,076
10000	10,478	10,775	10,612	0,055
100000	10,611	10,710	10,651	0,016

Taulukkoon 2.1 on koottu Monte Carlo integroimisen tuloksia, kun arvontapisteiden lukumäärää vaihdellaan. Testi suoritettiin 10 kertaa jokaisella eri arvontapisteiden lukumäärällä. Taulukosta huomataan, että mitä suurempi määrä pisteitä arvotaan, niin sitä vähemmän keskiarvo heittää oikeasta arvosta. Arvontapisteiden lukumäärän kasvattaminen myös pienentää saatujen integraaliarvojen varianssia, sillä 10:n toiston testissä pienimmän ja suurimman integraalin arvon erotus on hyvin pieni sekä 10000:tta että 100000:tta arvontapistettä käyttämällä. Tätä menetelmää käytettäessä tulee koordinaattipisteitä arpoa mahdollisimman paljon, jotta testi antaa tarkkoja likiarvoja integraaleille.

2.4 Monte Carlo -permutaatiotesti

Monte Carlo -permutaatiotesti on perusajatukseltaan hyvin yksinkertainen, sillä siinä toistetaan samaa toistoperiaatetta lukuisia kertoja. Käsiteltävä aineisto voidaan permutaatiotestissä jakaa kahteen osaan. Ensimmäisen osa on vertailuryhmä, jota merkitään tässä A:lla. Permutaatiotestiä tehdessä tutkitaan joukon A alkioiden arvoja ja verrataan niitä koko otospopulaatioon. Olkoon otospopulaatio B ja se sisältää kaikki permutaatiotestissä mukana olevat alkiot eli siten $A \subset B$. Permutaatiotestissä lasketaan ensiksi tutkittavat arvot vertailuryhmälle A ja tämän jälkeen tallennetaan saadut tiedot ylös. Seuraavaksi arvotaan vertailuryhmän alkioiden lukumäärän verran alkioita koko populaatiosta B ja lasketaan tästä joukosta samat tutkittavat asiat kuin vertailuryhmästä A. Tutkittavina asioina voivat olla esimerkiksi otoskeskiarvo, otosvarianssi tai otosihajonta.

Permutaatiotestin mallintamisen idea liittyy siihen, että tehdään tehdään nollahypoteesi H_0, jonka mukaan sekä vertailu että otospopulaation arvot ovat samansuuruisia. Tämän jälkeen tehdään H_1 hypoteesi, joka on nollahypoteesia vastaan ja tutkitaan sitä, että voidaanko mallintamisen yhteydessä saatujen tulosten perusteella hylätä nollahypoteesi H_0. Vaikka permutaatiotestillä voidaan käsitellä erityyppisiä aineistoja tehokkaasti, niin aineiston testaamiseen liittyy kuitenkin monia käytännön ongelmia. Näistä kerrotaan tarkemmin luvussa 3, jossa permutaatiotestiin liittyviä ominaisuuksia käsitellään yksityiskohtaisesti.

3. TUNTEMATTOMAN AINEISTON TUTKIMINEN MONTE CARLO -PERMUTAATIOTESTILLÄ

Tässä luvussa tutustutaan Monte Carlo -permutaatiotestin perusperiaatteeseen sekä tutkitaan permutaatiotestin lopputulokseen vaikuttavia tekijöitä. Permutaatiotestissä hyödynnetään paljon todennäköisyysmatematiikkaa permutaatioiden ja kombinaatioiden muodossa sekä hyödynnetään kertoman käsitettä. Kappaleessa tutkitaan myös permutaatiotestin hyviä sekä huonoja puolia. Samalla pohditaan permutaatiotestin hyödyntämisen mahdollisuutta ja järkevyyttä erikokoisten aineistojen käsittelyssä.

3.1 Permutaatiotestin toimintaperiaate ja siinä käytettävä matematiikka

Permutaatiotestillä voidaan tutkia tehokkaasti erikokoisia aineistoja luotettavasti, kun ymmärtetään permutaatiotestin lopputulokseen vaikuttavat tekijät. Olkoot vertailuryhmän alkioiden joukko A ja kaikki alkiot sisältävä joukko B. Koska joukko A käsittää vain osan joukon B alkioista, niin joukoille A ja B pätee $A \subset B$. Siten myös joukkojen A ja B sisältämien alkioiden lukumäärille $n(A)$ ja $n(B)$ pätee $n(A) < n(B)$, koska vertailuryhmän sisältäessä kaikki alkiot ei permutaatiotestissä ole mitään järkeä vertailuryhmän ollessa koko otospopulaation alkioiden joukko.

Tässä työssä kehitetään permutaatiotestistä tilastollista toistokoetta muistuttava mallinnusohjelma, jossa verrataan vertailuryhmää koko muuhun otospopulaatioon. Ennen testin suorittamista määritetään hypoteesit, joita testissä lähdetään tarkastelemaan. Yleensä hypoteesejä tehdään kaksi kappaletta. Ensimmäiseksi määritetään nollahypoteesi H_0 ja tämän jälkeen määritetään vaihtoehtoinen hypoteesi H_1, joka on nollahypoteesiä vastaan. Testin lähtöoletus on, että nollahypoteesi H_0 pitää paikkaansa [8, s. 29]. Kuitenkin testituloksesta riippuen nollahypoteesi H_0 voidaan hylätä ja todetaan, että vaihtoehtoinen hypoteesi H_1 pitääkin paikkaansa tehdyn testin perusteella. Jos H_1 hypoteesi osoittautuu oikeaksi yhden testin perusteella,

niin testi on hyvä tehdä uudestaan, koska satunnaisuudesta johtuen testi voi antaa virheellisen tuloksen jopa pienillä riskitasoilla. Testin luotettavuutta arvioidaan vielä tarkemmin luvussa 4.

Jos A on vertailuryhmän alkioiden joukko ja B on kaikki alkiot sisältävä joukko, niin hypoteesit permutaatiotestissä voidaan määritellä esimerkiksi seuraavasti.

$H_0 : \overline{X}_A = \overline{X}_i$ (nollahypoteesi)
$H_1 : \overline{X}_A \geq \overline{X}_i$ (ykköshypoteesi)

Hypoteeseissä $\overline{X}_A$ kuvaa vertailuryhmän alkioiden keskiarvoa ja $\overline{X}_i$ kuvaa koko otospopulaatiosta B otetun osajoukon keskiarvoa. Osajoukko $\overline{X}_i$ sisältää saman määrän alkioita kuin vertailuryhmä A. Permutaatiotestin lähtökohtana toimii oletus, että H_0 pitää paikkaansa. H_1 hypoteesi määritellään permutaatiotestissä tutkittavan asian mukaan. Tässä tapauksessa halutaan tietää, että onko vertailuryhmän alkioiden keskiarvo suurempi tai yhtä suuri, kuin satunnaisesti koko otospopulaatiosta valittujen alkioiden keskiarvo.

Testin lopputuloksen arvioimiseen käytetään yleensä P-arvoa. Hypoteesejä tutkivassa testissä P-arvo kertoo sen todennäköisyyden, jolla saatuun testitulokseen päästään käyttämällä H_0 hypoteesia (olettaen, että H_0 pitää paikkaansa). Olkoon $\overline{X}_i$ otospopulaation B osajoukko, jossa on alkioita yhtä paljon kuin joukossa A. Tällöin voidaan merkitä niitä $\overline{X}_i$ osajoukkoja x:llä, jotka toteuttavat ehdon $\overline{X}_A < \overline{X}_i$. Edellä määritettyjen hypoteesien perusteella P-arvo saadaan tässä tapauksessa jakamalla x toistomäärällä n. Jos P-arvo on hyvin matala, niin saatuun tulokseen päästään harvoin käyttämällä H_0 hypoteesiä. Tässä tapauksessa on syytä pohtia H_0:n hylkäämistä. Koska tilastolliseissa testeissä tutkitaan satunnaissuutta eri tavoilla, niin siten myös P-arvo on satunnaisluku ja sen suuruus vaihtelee toistettaessa testi täysin samalla tavalla. [8, s. 34–35] Permutaatiotestissä voidaan valita riskitaso α, jonka perusteella saadaan ehto H_0 hypoteesin hylkäämiselle. Mikäli riskitaso α on määritetty, niin H_0 hypoteesi hylätään, jos toteutuu ehto P-arvo $< \alpha$. Täten α kertoo sen todennäköisyyden, että hypoteesi hylätään virheellisesti saadun testituloksen perusteella.

Ennen permutaatiotestin suorittamista määritellään testissä tutkittavat hypoteesit H_0 ja H_1 sekä testin käyttäjän halutessa päätetään riskitaso α. Permutaatiotestin suorittavaa mallinnusohjelmaa varten käyttäjän tarvitsee laskea valmiiksi vertailuryhmän keskiarvo ja sekä vetailuryhmän sisältämän alkioiden lukumäärä. Tämän jälkeen testin suoritus voidaan aloittaa syöttämällä laskettu keskiarvo ja alkioiden lukumäärä mallinnusohjelmaan. Myös testin toistomäärä N syötetään funktion pa-

rametreihin. Tämän jälkeen mallinnusohjelma arpoo N-kertaa vertailuryhmässä olevien alkioiden lukumäärän verran alkioita koko populaatiosta ja laskee keskiarvon jokaiselle erikseen arvotulle joukolle. Samaa alkiota ei voida valita kahdesti samalla toistokerralla. Saatua keskiarvoa verrataan vertailuryhmän keskiarvoon ja lopulta N-kertaa ohjelman suoritettua mallinnusohjelma antaa käyttäjälle P-arvon, joka kertoo sen todennäköisyyden, jolla samaan tilanteeseen päästäisiin käyttämällä H_0 hypoteesiä.

Monte Carlo -permutaatiotestiin liittyy olennaisesti se, että monellako eri tavalla otospopulaation alkiot voidaan järjestää, ja montako erilaista x-lukumäärän sisältävää alkion yhdistelmää testissä on mahdollista arpoa. Tämän ongelman ratkaisemiseksi määritellään kertoma seuraavasti [9, s. 49].

Määritelmä 6 . *Olkoon X ei-negatiivinen kokonaisluku. Tällöin luvun X kertomaa merkitään $X!$ ja kertoma ilmaistaan muodossa*

$$X! = \begin{cases} X \cdot (X-1) \cdot \ldots \cdot 1 & \textit{kun } X \geq 1 \\ 1 & \textit{jos } X = 0 \end{cases}$$

Kertoman avulla voidaan ilmaista kuinka monella eri tavalla voidaan joukon alkiot asettaa jonoon. Esimerkiksi joukko, joka sisältää n alkiota, voidaan järjestää $n!$ eri tavalla jonoon [9, s. 49].

Permutaatiotestissä ei olla kuitenkaan kiinnostuneita siitä, että kuinka monella eri tavalla joukon alkiot voidaan asettaa jonoksi. Olennaista on tietää montako erilaista k-alkioista osajoukkoa voidaan valita x-alkioisesta joukosta. Osajoukkojen lukumäärä saadaan selville hyödyntämällä kertoman määritelmää seuraavasti [9, s. 58-59].

Määritelmä 7 *Joukosta, jossa on x alkiota, voidaan valita k-alkioisia osajoukkoja $\binom{x}{k}$ kappaletta, jossa*

$$\binom{x}{k} = \frac{x!}{k!(n-k)!}$$

Tässä tapauksessa k-alkioisia osajoukkoja sanotaan joukon k-kombinaatioiksi. Kombinaatioiden määrä on tärkeä osa permutaatiotestiä, sillä arvottaessa vertailuryhmän A alkioiden lukumäärän verran alkioita koko otospopulaatiosta B saadaan mahdollisten kombinaatioiden määräksi $\binom{B}{L}$, missä L kertoo vertailuryhmän alkioiden

lukumäärän. Permutaatiotestistä voisi saada mahdollisimman tarkan arvion käymällä kaikki mahdolliset kombinaatiot läpi tasan kerran, ja laskemalla siitä, että kuinka suuri osa kombinaatioista toteuttaa H_1 hypoteesin. Otoskoon suurentuessa tarpeeksi kaikkien kombinaatioiden läpikäyminen ei kuitenkaan ole järkevää tietokoneen laskentatehon puitteissa, joten siksi riittävän monta toistoa suorittava mallinnusohjelma pystyy antamaan luotettavia tuloksia suurten otospopulaatioiden tapauksissa, kunhan toistomäärä N on riittävä.

3.2 Normaalijakauman hyödyntäminen permutaatiotestissä

Monte Carlo -permutaatiotestin luotettavuutta oin vaikea arvioida, sillä suurimmassa osassa tapauksissa otospopulaatio on riittävän suuri, jotta testin toistomäärä N ei riitä kaikkien mahdollisten kombinaatioiden läpikäymiseen, vaikka alkioiden arpomisen sijaan käytäisiin jokainen mahdollinen kombinaatio tasan kerran lävitse. Kuinka sitten asiaa voidaan tarkastella? Määritetään aluksi hypotessit seuraavasti.

$H_0 : \overline{X}_A = \overline{X}_i$
$H_1 : \overline{X}_A > \overline{X}_i$

Nyt asian yksinkertaistamiseksi merkitään niitä kombinaatioita arvolla 1, jotka totettuvat ehdon $\overline{X}_A \leq \overline{X}_i$ ja niitä arvolla 0, jotka toteuttavat ehdon $\overline{X}_A > \overline{X}_i$. Tässä $\overline{X}_i$ on permutaatiotestin arpoman joukon alkioiden keskiarvo ja $\overline{X}_A$ vertailuryhmän keskiarvo.

Tällä tavalla jokainen kombinaatio saa arvon 0 tai 1. Kombinaatiot, jotka saavat arvon 1 ovat H_1 hypoteesin kannalta epäedullisia ja arvon 0 omaavat kombinaatiot ovat suotuisia H_1 näkökulmasta. Koska P-arvo kertoo testissä olevan H_0 hypoteesin paikkaansa pitävyyden todennäköisyyden ja tässä tapauksessa samalla H_1 hypoteesin hylkäämisen todennäköisyyden, niin laskettaessa testauksen kombinaatioiden arvojen (arvot ovat 0 tai 1) odotusarvo saadaan samalla selvitettyä P-arvo.

Kun merkitään kombinaatioiden arvoja äskeisten ehtojen perusteella joko luvulla 0 tai 1, ja toistetaan testiä N-kertaa, niin saadaan luotua toistokoe, jossa on kaksi mahdollista vaihtoehtoa. Toistokokeen sisältäessä riittävän monta toistoa voidaan normaalijakaumaa hyödyntää testin tuloksen ennustamisessa, jos kombinaatioiden saamien arvojen odotusarvo ja varianssi tiedetään.

Normaalijakauman avulla voidaan tutkia myös muiden Monte Carlo -menetelmien tarkkuutta. Kuvassa 2.1 esitettiin MC -integroimisen tuloksia, kun hyödynnetään satunnaislukujen arvonta -menetelmää. Kyseisen menetelmän tarkkuutta voidaan

yhtä lailla arvioida normaalijakaumaa käyttämällä, koska merkitsemällä integraalin sisään osuvia pisteitä arvolla 1 ja integraalin ulkopuolella olevia pisteitä arvolla 0, päädytään kaksi vaihtoehtoa sisältävään toistokokeeseen aivan kuten permutaatiotestissä. Tämä on hyvä esimerkki siitä, että yllättävän moni matemaattinen numeerisen ratkaisun antava menetelmä palautuu pienellä muotoilulla tilanteeseen, jossa normaalijakaumaa pystytään hyödyntämään tehokkaasti. Siksi normaalijakauma on tärkeä työkalu monissa tilastollisissa kokeissa. Määritellään normaalijakauma seuraavasti [5, s. 45–47].

Määritelmä 8 *Olkoot X satunnaismuuttuja ja S_X satunnaismuuttujan otosavaruus. Satunnaismuuttuja X on normaalijakautunut parametrein (μ, σ^2), jos $S_X \in \mathbb{R}$ ja X:n tiheysfunktio $f(x)$ on muotoa*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma}^2)}$$

Tällöin satunnaismuuttuja X on normaalijakautunut ja merkitään $X \sim N(\mu, \sigma^2)$. Huomataan myös, että eksponentissa oleva lauseke $(\frac{x-\mu}{\sigma})^2$ saa neliöön korotuksen johdosta saman arvon osoittajan merkistä riippumatta. Koska x:ää lukuunottamatta muut muuttujat pysyvät vakioina, niin tästä seuraa, että $f(x)$ on symmetrinen suoran $x = \mu$ suhteen. Toinen tärkeä ominaisuus normaalijakaumassa on sen standardoitu muoto, jonka avulla voidaan laskea helposti kertymäfunktion arvo. kertymäfunktion likiarvo pystytään katsomaan myös helposti taulukosta.

Jos muuttuja Z on standardoidusti normaalijakautunut, niin sitä voidaan merkitä $Z \sim N(0, 1)$. Standardoidun normaalijakautuneen muuttujan Z kertymäfunktio voidaan kirjoittaa seuraavasti [5, s. 45].

$$\phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} dx \tag{3.1}$$

Todistus sivuutetaan sen haastavuuden vuoksi. Jotta pystytään hyödyntämään standardoidusti normaalijakautuneen muuttujan kertymäfunktiota, niin normaalijakautunut satunnaismuuttuja $X \sim N(\mu, \sigma^2)$ muutetaan käsiteltävään muotoon määritelmän 9 perusteella.

Määritelmä 9 *Jos X noudattaa normaalijakaumaa eli pätee $X \sim N(\mu, \sigma^2)$, niin silloin*

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

Määritelmän 9 avulla voimme muuttaa minkä tahansa normaalijakaumaa noudattavan satunnaismuuttujan standardoituun muotoon riippumatta μ ja σ^2 arvosta. Nyt päästään käsiksi satunnaismuuttujan kertymäfunktioon käyttämällä seuraavaa lausetta apuna [5, s. 46].

Lause 3 *Jos satunnaismuuttuja X noudattaa normaalijakaumaa $X \sim N(\mu, \sigma^2)$, niin X:n kertymäfunktio $F(x)$ on silloin*

$F(x) = \phi(\frac{x-\mu}{\sigma})$.

Todistus. Hyödynnetään määritelmää 9, jotta kertymäfunktio voidaan muuttaa sellaiseen muotoon, josta voidaan laskea kertymäfunktion arvoja käyttämälllä standardoidun normaalijakaumaa hyödyksi. Kertymäfunktioksi saadaan siis

$$F(x) = P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) \overset{\text{Määr } 9}{=} P(Z \leq \frac{x-\mu}{\sigma}) = \phi(\frac{x-\mu}{\sigma}) = \phi(z) \quad \square$$

Permutaatiotestissä P-arvo lasketaan vertailemalla satunnaisesti arvottujen alkiojoukkojen keskiarvoja vertailuryhmän keskiarvoon. Tämän seurauksena P-arvo on satunnaismuuttuja, koska sen arvo riippuu arvotuista alkiojoukoista. Jotta permutaatiotestin avulla saadaan muodostettua mahdollisimman tarkka arvio P-arvon käyttäytymisestä, niin permutaatiotestin kombinaatioiden arvojen (0 tai 1) odotusarvo selvitetään ensin. Pienillä otospopulaatioilla kyseinen odotusarvo voidaan laskea käymällä jokainen mahdollinen kombinaatio läpi tasan kerran ja laskemalla odotusarvo siitä. Keskisuurilla tai suurilla otospopulaatioilla odotusarvoa ei voida lasketa tarkasti, sillä mahdollisten kombinaatioiden määrä kasvaa niin paljon, että ei ole järkevää edes tietokoneella käydä kaikkia mahdollisia kombinaatioita lävitse. Normaalijakaumaa pystytään hyödyntämään vain, jos odotusarvon lisäksi jakauman varianssi tiedetään. Yksittäisen toiston varianssi sellaisessa toistokokeessa, joissa on mahdollista saada kaksi eri vaihtoehtoa (tässä kombinaation arvot 0 tai 1), saadaan lauseesta 4 ratkaistua odotusarvon avulla seuraavasti

Lause 4 . *Olkoot X diskreetti satunnaismuuttuja ja μ satunnaismuuttujan odotusarvo. Jos satunnaismuuttuja X voi saada vain arvot 0 tai 1, niin satunnaismuuttujan varianssi voidaan ratkaista yhtälöstä*

$$\sigma^2 = -\mu^2 + \mu$$

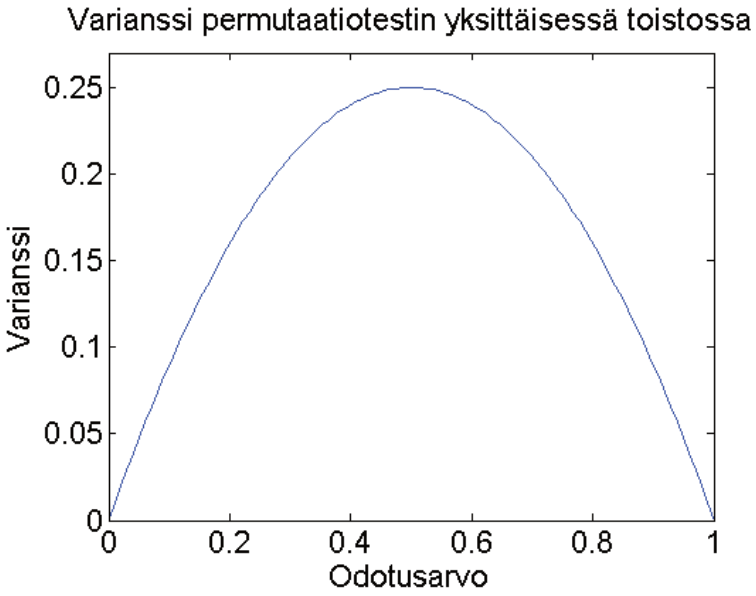
Todistus. Käytetään apuna määritelmää 3, jonka mukaan varianssi voidaan ilmoittaa muodossa $\sigma^2 = \sum p_i(x_i - \mu)^2$. Olkoon x se todennäköisyys, että kombinaation arvoksi tulee 1. Siten $(1-x)$ kuvaa sitä todennäköisyyttä, että kombinaation arvoksi tulee 0. Nyt määritelmän 3 lauseke saadaan muotoon

$$\begin{aligned} & \sigma^2 &=& \sum p_i(x_i - \mu)^2 \\ \Leftrightarrow\ & \sigma^2 &=& x(1-\mu)^2 + (1-x)(0-\mu)^2 \\ \Leftrightarrow\ & \sigma^2 &=& x(1-2\mu+\mu^2) + (1-x)\mu^2 \\ \Leftrightarrow\ & \sigma^2 &=& x - 2x\mu + x\mu^2 + \mu^2 - x\mu^2 \\ \Leftrightarrow\ & \sigma^2 &=& x - 2x\mu + \mu^2 \end{aligned}$$

Voidaan sijoittaa $x = \mu$, koska x ilmaisee todennäköisyyttä, että kombinaation arvoksi tulee 1 eli samalla x ilmaisee myös P-arvon, joka kuvastaa kombinaatioiden arvojen odotusarvoa μ.

$$\begin{aligned} \Leftrightarrow\ & \sigma^2 &=& \mu - 2\mu\mu + \mu^2 \\ \Leftrightarrow\ & \sigma^2 &=& -\mu^2 + \mu \qquad \square \end{aligned}$$

Piirretään yhtälöstä $-\mu^2 + \mu$ kuvaaja, jonka avulla voidaan tutkia varianssin vaihtelevuutta eri odotusarvoilla.



Kuva 3.1 *Satunnaismuuttujan varianssin suuruus yksittäisessä toistossa*

Kuten kuvasta 3.1 voidaan selvästi huomata, niin varianssi saa suurimman arvon funktion $f(x) = -\mu^2 + \mu$ derivaatan nollakohdassa $\frac{1}{2}$, joten varianssin suurin mahdollinen arvo on $f(\frac{1}{2}) = \frac{1}{4} = 0,25$. Alaspäin aukeavana paraabelina varianssi pienenee odotusarvon lähestyessä arvoa 0 tai 1. Koska hypoteesejä testatessa H_0 hypoteesi voidaan hylätä P-arvon (eli μ:n) lähestyessä arvoa 0 tai 1 riippuen hypoteesien määrittelytavasta, niin yksittäisen toiston varianssi on yleensä paljon pienempi kuin $\mu_{max} = 0,25$. Kun yksittäisen toiston varianssi pienenee, niin myös testin kokonaisvarianssi pienenee, mikä parantaa entisestään permutaatiotestin luotettavuutta.

Yksittäisen toiston varianssia ei kuitenkaan voida suoraan käyttää P-arvon arvioimiseen, vaan P-arvon arvioimiseksi tarvitsee selvittää ensiksi N kertaa toistettavan testin kokonaisvarianssi. Tähän päästään käsiksi keskeisen raja-arvolauseen avulla.

Lause 5 . ***Keskeinen raja-arvolause.*** *Olkoon X satunnaismuuttuja, jonka odotusarvo on μ ja varianssi σ^2. Jos $X_1, X_2, \ldots, X_n$ on satunnaismuuttujasta X otettu otos ja n on otosalkioiden lukumäärä, niin standardoidun otoskeskiarvon $\overline{X}$ kertymäfunktio $F(t)$ lähestyy $N \sim (0,1)$-jakauman kertymäfunktiota $\phi(z)$ eli*

$$F(t) = P(\overline{X} \leq t) \to \phi(z), \quad kun \;\; n \to \infty$$

Lause on 5 tärkeä, sillä satunnaismuuttujien summan jakauma alkaa lähestymään normaalijakaumaa, kun summaan laskettavia satunnaismuuttujan arvoja on paljon. Kun approksimoidaan satunnaismuuttujan summaa normaalijakaumalla, saadaan useimmiten hyviä arvioida toistomäärän ollessa yli 30, vaikka satunnaismuuttujan jakaumaa ei tunneta. [5, s. 66–67] Lauseen 5 perusteella satunnaismuuttujan X summalle saadaan seuraava ominaisuus.

Lause 6 . *Olkoon $X_1 + X_2 + \ldots + X_n$ satunnaismuuttujan X otos. Jos satunnaismuuttujan odotusarvo on μ, varianssi σ^2 ja satunnaismuuttujan alkioiden lukumäärä n, niin tällöin satunnaismuuttujan summalle S ja otoskeskiarvolle $\overline{X}$ saadaan*

$$S = X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$$

$$\overline{X} \sim N(\mu, \tfrac{\sigma^2}{n})$$

Permutaatiotestin ohjelma arpoo otospopulaatiosta alkiojoukon, jolle se antaa arvon 0 tai 1 riippuen siitä, että toteuttaako alkiojoukko H_1 hypoteesin. Täten ohjelma toimii ikään kuin satunnaislukujen arpomiskoneena, koska se arpoo kombinaation,

joka saa arvon 0 tai 1. Lauseen 6 perusteella tämä tarkoittaa samalla myös sitä, että kombinaatioiden arvojen summa lähestyy normaalijakaumaa, kun toistomäärä N suurenee. Saatu summa voidaan muuttaa standardoituun muotoon määritelmän 9 avulla, jos tiedetään kombinaatioiden arvojen odotusarvo ja varianssi. Ilman erillisiä tietokoneohjelmia vain harvassa tapauksessa odotusarvo tiedetään (otosvarianssi ja otoskeskiarvo eivät kelpaa), jolloin standardoidun normaalijakauman kertymäfunktiota ei voida hyödyntää kaavan 3.1 mukaisesti kuin todella harvoissa tapauksissa.

3.3 Permutaatiotestin hyvät ja huonot ominaisuudet

Monte Carlo -permutaatiotestillä on sekä hyviä että huonoja puolia. Permutaatiotestin vahvuuksiin kuuluu ehdottomasti se, että riittävän suurilla toistomäärillä jopa isoista aineistoista saadaan kelvollisia likiarvoja P-arvoille. Vahvuuksiin kuuluvat myös testin yksinkertainen toimintaperiaate, joka auttaa ymmärtämään testin prosessia ja siihen vaikuttavia asioita sekä tietokoneen nopea laskentatehokkuus, sillä yksinkertaisen perusrakenteen vuoksi testi on usein kevyt suorittaa. Tämän seurauksena permutaatiotesti suurilla toistomäärillä saadaan suoritettua hyvin lyhyessä ajassa, joten testin uudelleensuorittaminen tarvittaessa onnistuu vaivattomasti, ja itse testin luotettavuuden arviointiin jää enemmän aikaa, kun testin suorittamiseen ei kulu ylimääräistä aikaa.

Permutaatiotestin heikkoudet liittyvät pitkälti testin tuloksen analysoimiseen sekä suurten otospopulaatioiden tuloksien luotettavuuteen. Ongelmia syntyy nimenomaan suurten aineistojen tutkimisessa, sillä otospopulaation kasvaessa mahdollisten kombinaatioiden määrä kasvaa eksponentiaalisesti. Pienillä ja keskikokoisilla otospopulaatioilla ongelmaa ei vielä ole kombinaatiomäärän pysyessä inhimillisenä. Jo hyvin pienillä joukkojen A ja B alkioiden lukumäärillä on vaikea laskea kaikkia mahdollisia tapauksia [10, s. 99–100]. Tämä huomataan helposti jo melko pienillä otospopulaatioilla, sillä esimerkiksi $\binom{10}{5} = 252$, $\binom{20}{5} = 15504$ ja $\binom{50}{5} = 2118760$. Toisaalta kombinaatioiden suuri määrä ei ole niin suuri ongelma, sillä suuri toistomäärä auttaa luotettavan tuloksen saamisessa, vaikka kombinaatioita olisikin reilusti enemmän kuin toistoja testissä. Toistokokeen omaisena kokeena myös normaalijakaumaa pystytään hyödyntämään testin analysoimisessa, jos odotusarvo ja varianssi tunnetaan. Tarkempia tulkintoja testin suorittamiseen ja tulkintaan liittyen on esitetty luvussa 4.

4. PERMUTAATIOTESTIN MALLINTAMINEN

Tässä kappaleessa perehdytään permutaatiotestin mallintamiseen liittyviin asioihin. Ensinki selvitetään kuinka tiedot populaation arvoista luetaan mallinnusohjelmaan ja minkämuotoista tietoa mallinnusohjelma tarvitsee toimiakseen. Myös permutaatiotestin mallintamisessa käytettävät merkinnät selvitetään, jotta käyttäjä pystyy laskemaan ja syöttämään tarpeelliset alkutiedot ohjelmaan. Lisäksi tutkitaan testin lopputulokseen vaikuttavien asioiden riippuvuutta ja selvitetään keinoja, joilla virhettä pystytään pienentämään järkevillä keinoilla.

4.1 Mallin toiminta ja siinä käytetyt merkinnät

Permutaatiotestiä hyödynnettäessä P-arvon selvittämiseksi käyttäjältä vaaditaan ennen varsinaisen testin suorittamista sekä vertailuryhmän alkioiden keskiarvon, että vertailuryhmässä olevien alkioiden lukumäärän laskemista. Testin onnistumisen kannalta on erittäin tärkeää laskea oikein keskiarvo ja alkioiden lukumäärä, sillä molemmat vaikuttavat merkittävästi lopputulokseen. Kun nämä kaksi asiaa tiedetään, voidaan testi aloittaa syöttämällä vertailuryhmän keskiarvo, sen alkioiden lukumäärä sekä toistojen määrä permutaatiotestissä.

Permutaatiotestissä vertailuryhmän keskiarvoa, jonka käyttäjä syöttää mallinnusohjelmaan, merkitään K:lla ja puolestaa vertailuryhmän sisältävien alkioiden lukumäärää L:llä. Mallinnusohjelmassa L1:llä merkitään koko otospopulaation käsittävän joukon alkioiden lukumäärää. Testin käyttäjän tulee huomioida, että funktion parametriksi tulee syöttää nimenomaan L:än arvo eli vertailuryhmän alkioiden lukumäärä eikä koko otospopulaation alkioiden lukumäärää. Mallinnusohjelma pystyy itse automaattisesti laskemaan L1 suuruuden, kun ohjelma lukee tiedot käyttäjän valitsemasta tiedostosta. Parametreihin syötetään vielä tieto toistomäärästä, jota merkitään normaaliin tapaan N:llä.

Käyttäjän halutessa suorittaa permutaatiotestin avataan Liitteen 1 mukainen Matlab-ohjelma, ja syötetään Matlabin komentoriville alkutiedot seuraavassa muodossa.

permutaatiotesti(K, L, N);

K:n paikalle syötetään vertailuryhmän keskiarvo, L:n paikalle vertailuryhmän alkioiden lukumäärä ja N:n paikalle testin toistomäärä. Desimaalilukuja kirjoittaessa tarkistetaan, että Matlabiin on syötetty pilkun sijaan piste, koska Matlab ei tunnista desimaalipilkkua. Alkutietojen jälkeen kysytään testissä käytettävää tiedostoa, josta saadaan alkioiden arvot permutaatiotestiä vaten. Tiedoston nimi syötetään ohjelmalla kokonaisuudessaan, esim. data.txt, jonka jälkeen testi laskee P-arvon, jonka jälkeen ohjelma tulostaa sen näytölle automaattisesti.

Alla olevasta listasta nähdään kaikki oleelliset merkinnät permutaatiotestin mallinnusohjelmassa.

B = otospopulaation vaakavektori
E = H_1 hypoteesin toteuttavien arvontojen lukumäärä
K = vertailuryhmän alkioiden keskiarvo
L = vertailuryhmän alkioiden lukumäärä
L_1 = otospopulaation alkioiden lukumäärä
N = mallin suorittama toistomäärä
P-arvo = testauksen tuloksena saatava P-arvo
X = permutaatiotestin arpoman joukon keskiarvo (verrataan K:hon)

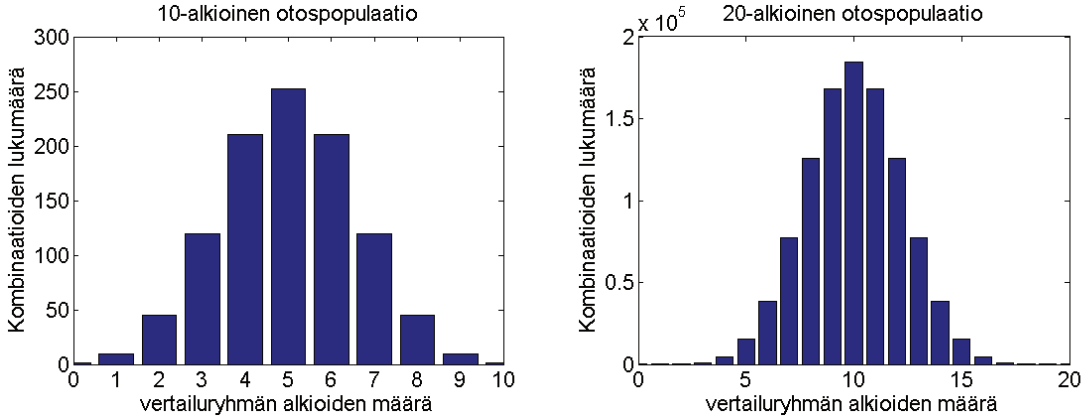
4.2 Aineiston muoto ja sen tuominen mallikoodiin

Permutaatiotestiä mallintava ohjelma tarvitsee oikeanlaisen tiedoston, jotta se pystyy käsittelemään saatua tietoa. Koska Matlabin ominaisuuksista johtuen siinä voi olla vain yksi funktio yhdessä ohjelmassa, niin se rajoittaa hieman tietojen tuomista mallinnusohjelmaan. Mallinnusohjelma voitaisiin laittaa käyttäjältä kysymään tarkasti, että mistä tiedostosta, miltä sarakkeelta, ja kuinka paljon tietoja ohjelmaan halutaan tuoda, mutta monen syötteen antaminen täsmällisesti yhtä testiä varten on työlästä. Lisäksi käyttäjän tekemien virheiden mahdollisuus kasvaa suuresti. Tämän takia liitteessä 1 esitetyllä mallinnusohjelman koodilla pystytään lukemaan tietoa syötetyn tiedoston ensimmäiseltä pystysarakkeelta, joissa kaikkien arvojen on oltava numeroita (käyttävä desimaalipistettä, koska Matlab ei ymmärrä desimaalipilkkua). Tämä on perusteltua jo pelkästään sillä, että testikäyttäjä joutuu laskemaan itse keskiarvon ja alkioiden lukumäärän, joten aineiston on järkevää olla yhdellä sarakkeella, sillä se helpottaa myös käyttäjän keskiarvon laskemista. Tämän seurauksena mallinnusohjelman toiminta saadaan pidettyä mahdollisimman yksinkertaisena kaikille käyttäjille, koska se ei vaadi käyttäjältä kuin alkusyötteiden ja tiedoston nimen syöttämisen testin suorittamiseksi.

Mallinnusohjelma vaatii myös tietyn tyyppisen tiedstomuodon toimiakseen. Liitteen 1 malli pystyy lukemaan tietoa joko txt- tai csv-tiedostosta. Näillä tiedostomuodoilla testi toimii luotettavasti. Monissa tilastollisissa ohjelmissa käytetään csv-tiedostoja aineistojen tutkimisessa, joten siksi permutaatiotestin toimiminen csv-tiedostolla voi helpottaa testaajan työtä huomattavasti, jos käsiteltävä aineisto on valmiiksi oikeassa muodossa. Testi toimii myös txt-tiedostoilla, joten pieniä aineistoja tutkittaessa käyttäjä pystyy tarvittaessa syöttämään tiedot itse tekstitiedostoon. Tämän seurauksena permutaatiotestin käyttäminen aineistojen tutkimuksessa on helppoa.

4.3 Vertailuryhmän ja otospopulaation vaikutus kombinaatioiden määrään

Vertailuryhmän A ja otospopulaation B koko vaikuttavat merkittävästi erilaisten kombinaatioiden lukumäärään. Kuten jo luvussa 3 mainittiin, pienillä otospopulaatioilla kombinaatioiden määrät pysyvät vielä inhimillisinä, mutta suurten otoskokojen tapauksissa ei pystytä käymään läpi kaikki mahdollisia kombinaatioita. Kuva 4.1 havainnollistaa kombinaatioiden määrää pienillä otospopulaatioilla.

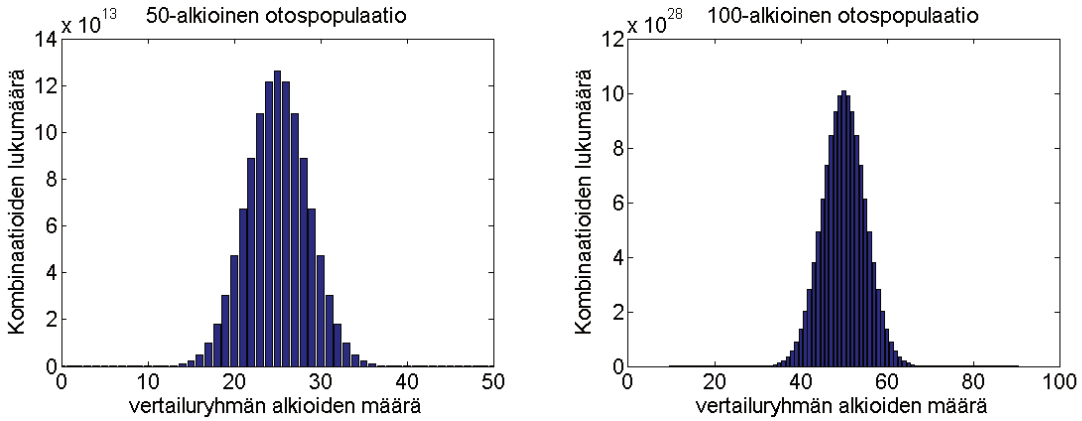


Kuva 4.1 *Mahdollisten kombinaatoiden määrä 10-alkioisella ja 20-alkioisella otospopulaatiolla*

Vasemmanpuoleinen pylväsdiagrammi ilmaisee tapauksen $\binom{10}{a}$ ja oikeanpuoleinen kuvaaja tapauksen $\binom{20}{a}$, joissa a kuvaa vertailuryhmässä olevien alkioiden lukumäärää. Kuvan 4.1 pylväsdiagrammeista huomataan, että jo pienikin otospopulaation kasvattaminen lisää kombinaatioiden määrän moninkertaiseksi. Lisäksi kuvista voidaan todeta myös, että kombinaatioiden määrä vertailuryhmän alkioiden funktiona on symmetrinen ja kombinaatioita on eniten, kun $n(a) = \frac{1}{2}n(b)$, jossa $n(a)$ on vertailuryhmän ja $n(b)$ otospopulaation alkioiden lukumäärä. Kuvasta voidaan myös päätellä, että permutaatiotestin mahdollisten kombinaatioiden lukumäärä pysyy siedettävänä, jos vertailuryhmän koko on pieni suhteessa otospopulaatioon.

Toisaalta pientenkin otospopulaatioiden tapauksissa permutaatiotestiin liittyy eräs suuri ongelma. Vaikka kombinaatioita olisikin vain vähän, niin tämä aiheuttaa silti ongelmia odotusarvon laskemisessa. Ilman odotusarvoa ei voida hyödyntää normaalijakaumaa ja sitä kautta P-arvon satunnaisuutta ei pystytä arvioimaan tarkasti. Jotta odotusarvo saadaan selville, tarvitaan erillinen laskentaohjelma selvittämään odotusarvo käymällä kaikki mahdolliset permutaatiot läpi tasan kerran ja tutkimalla mitkä annetuista tapauksista toteuttavat H_1 hypoteesin ja mitkä eivät.

Jos otospopulaatio on suuri, niin kombinaatioita on jo niin paljon, että edes erillisten tietokoneohjelmat laskentatehokkuus ei riitä kaikkien permutaatioiden läpikäymiseen ja siten odotusarvon laskemiseen. Tämä käy ilmi kuvasta 4.2.



Kuva 4.2 *Mahdollisten kombinaatoiden määrä 50-alkioisella ja 100-alkioisella otospopulaatiolla*

Kuten kuvan 4.2 diagrammeista käyi ilmi, niin jo 50-alkioisella otospopulaatiolla kombinaatioiden määrä on valtava, ja 100-alkioisella otospopulaatiossa kombinaatioita voi olla jo yli 10^{28} kappaletta. Kuvan 4.2 oikeanpuoleista pylväsdiagrammin muotoa katsomalla voisi virheellisesti olettaa, että kombinaatioiden määrä pysyy järkevänä, kun vertailuryhmän koko on todella pieni verrattuna koko otospopulaatioon tai lähes koko otospopulaation suuruinen. Näin ei kuitenkaan ole, sillä esimerkiksi $\binom{100}{15} \approx 10^{17}$ ja $\binom{100}{10} \approx 10^{13}$.

Ongelmaksi muodostuu siis kombinaatioiden laskeminen odotusarvon selvittämiseksi. Suurten otospopulaatioiden tapauksissa erillisellä laskentaohjelmillakaan ei saada ratkaistua odotusarvoa järkevässä ajassa. Lisäksi ei ole järkevää olettaa, että käyttäjällä on käytössä mitään erillistä toimivaa laskentaohjelmaa eikä ole hyvä lähtökohta, että permutaatiotestin suorittamiseksi tarvitaan useita testaus- tai mallinnusohjelmia. Kuinka sitten odotusarvoa ja sitä kautta P-arvon satunnaisuutta ja suuruutta voidaan arvioida luotettavasti? Onneksi permutaatiotesti on hyvin kevyt suorittaa, joten käyttäjä pystyy itse tutkimaan P-arvon käyttäytymistä toistaessa

testiä mahdollisimman monta kertaa. Jo muutaman testauksen perusteella käyttäjä pystyy päättelemään P-arvon käyttäytymistä, sillä käytettäessä testissä suurta toistomäärää testin antaman P-arvon varianssi pysyy pienenä.

4.4 Virheen minimoiminen permutaatiotestissä

Moni tekijä aiheuttaa virhettä permutaatiotestin lopputulokseen. Vaikka työssä ei käsitelläkään satunnaislukujen generointia, niin testausohjelman kyky arpoa satunnaislukuja riittävän hyvin voi vaikuttaa lopputulokseen yllättävän paljon. Toistoja permutaatiotestissä voi olla jopa satoja tuhansia ja yksittäisessä toistossa ohjelma saattaa joutua arpomaan kymmeniä satunnaislukuja arpoessaan alkioita. Vaikka virhe algoritmissa ei olisikaan kovin suuri, niin sen vaikutus helposti moninkertaistuu toistomäärän ollessa suuri. Käyttäjä pystyy myös itse vaikuttamaan permutaatiotestin tarkkuuteen, sillä valitsemalla testiä suoritettaessa riittävän suuri toistomäärä n, saadaan testin antaman P-arvon varianssia pienennettyä.

Jos kombinaatioiden arvojen odotusarvo on pystytty ratkaisemaan esimerkiksi toisella mallinnusohjelmalla käymällä kaikki mahdolliset kombinaatiot läpi tasan kerran, niin P-arvon luottamusväli voidaan selvittää normaalijakauman avulla. Nyt voidaan lauseen 4 avulla laskea yksittäisen toiston varianssi. Tässä tapauksessa käyttämällä yksittäisen toiston maksimivarianssia $\mu_{max} = 0,25$ lauseen 4 avulla lasketun varianssin sijasta, saadaan luottamusväli kattamaan suurempi todennäköisyys. Mikäli P-arvon luottamusväli kattaa suuremman todennäköisyyden kuin sen pitäisi, niin hypoteesin hylkääminen, toteaminen oikeaksi tai hypoteesin epävarmuus on helpompi todeta. Toki maksimivarianssia käytettäessä käyttäjä ei saa täysin oikeanlaista kuvaa saadun P-arvon luottamusvälistä, mutta yleisellä tasolla permutaatiotesti antaa todella tarkkoja arvioita aineiston muodosta ja koosta riippumatta. Kaiken lisäksi se on kevyt suorittaa tietokoneella, mikä parantaa entisestään sen käyttämisen kannattavuutta.

Permutaatiotestiä tehdessä pitää kuitenkin muistaa, että keskiarvo ei välttämättä kuvaa parhaiten aineistoa. keskiarvo ei kuvaa ainestoa hyvin, jos aineistossa on esimerkiksi muutama todella suuri luku. Tämän seurauksena vertailuryhmän sisältäessä yhdenkin vähintään kertaluokan suuremman alkion voi olla lähes mahdotonta ottaa sellaisia alkiokombinaatioita, joilla olisi suurempi keskiarvo kuin vertailuryhmällä. Ongelmia voi muodostua myös siitä, jos alkioiden arvot painottuvat todella paljon ääripäihin. Ongelmaa korostaa vielä mahdollinen suuri ero pienimmän ja suurimman alkion välillä.

Ongelmatapauksissa käyttäjällä on mahdollisuus pilkkoa otospopulaatio osiin skaa-

laamalla alkioiden arvot sopivasti. Esimerkiksi otospopulaation sisältäessä arvoja väliltä [10, 10000] voi olla esimerkiksi järkevää skaalata arvot muutamaan luokkaan seuraavasti [10-50] = 0, [50-200] = 1, [200-1000] = 2 ja [1000-10000] = 3. Nyt arvot vaihtelevat 0-3 väliltä ja permutaatiotestillä voidaan saada jotain järkeviä tuloksia aineistosta. Näin tehdessä pitää kuitenkin muistaa, että arvoluokkien skaalaaminen vaikuttaa merkittävästi lopputuloksiin, joten tämänkaltaisia muutoksia tehdessä tuloksiin on hyvä suhtautua kriittisesti. Paljon ääripään arvoja tai yksittäisiä suuria arvoja sisältäviä aineistoja on vaikea tutkia millään keinolla, mutta permutaatiotesti antaa siihen kuitenkin mahdollisuuden. Tuloksiin on kuitenkin hyvä suhtautua kriittisesti, mikäli aineistoa skaalataan erikseen permutaatiotestiin.

5. JOHTOPÄÄTÖKSET

Tässä työssä tutustuttiin pintapuolisesti erilaisiiin Monte Carlo -menetelmiin ja työn pääpaino oli Monte Carlo permutaatiotestissä. Monia MC -menetelmiä pystytään hyödyntämään monipuolisesti ja useimmat menetelmät antavat hyviä likiarvoja ongelmiin, joiden ratkaiseminen tarkasti voi olla hyvinkin hankalaa. Usein hyvä likiarvo on riittävä ongelman ratkaisemiseksi. Tyypillistä erilaisille Monte Carlo -menetelmille on se, että menetelmät käyttävät usein hyvin yksinkertaista matematiikkaa. Menetelmien tehokkuus ja käytettävyys perustuukin niiden helppoon mallintamiseen sekä yksinkertaisuuteen. Mallinnusohjelmat ovat usein kevyitä suorittaa tietokoneella, jolloin toistomääriä saadaan kasvatettua suuriksi, mikä puolestaan parantaa entisestään testien tarkkuutta.

Monte Carlo -menetelmien keskeinen ongelma liittyy kuitenkin satunnaislukujen arpomiseen. Satunnaislukugeneraattorit ja algoritmit eivät koskaan pysty arpomaan täysin satunnaisia lukuja, joten ohjelman satunnaislukujen arpomisen tarkkuus voi vaikuttaa testin lopputulokseen. Useimmat ohjelmat kuitenkin pystyvät arpomaan satunnaislukuja riittävän hyvin, minkä takia eri testeillä ja metodeilla saadut likiarvot ovat usein riittävän tarkkoja.

Yllättävän monessa menetelmässä pystytään hyödyntämään normaalijakaumaa jollain tavalla testituloksien analysoimisessa. Normaalijakaumaa voidaan hyödyntää niin MC -integroimista satunnaislukujen arvonta -menetelmää käytettäessä kuin MC -permutaatiotestiä käytettäessä. Menetelmissä voidaan pienillä merkitsemistavoilla palauttaa ne muotoon, jossa normaalijakaumaa voidaan hyödyntää testituloksien arvioimisessa. Merkitsemällä permutaatiotestissä kaikki vaihtoehtoisen hypoteesin H_1 kannalta suotuisat kombinaatiot arvolla 0 ja epäsuotuisat arvolla 1 päädytään tilanteeseen, jossa kombinaatioiden odotusarvo kertoo samalla P-arvon, joka ilmaisee todennäköisyyden, että testin tulokseen olisi päästy käyttämällä H_0 hypoteesiä.

Permutaatiotesti on luotettava, koska merkitsemällä kombinaatioiden arvoja luvuilla 0 ja 1, saadaan luotua toistokoe, jossa on vain kaksi mahdollista vaihtoehtoa. Koska tietokoneohjelma pystyy lisäksi laskemaan tuhansia toistoja todella nopeassa ajassa, niin testin lopputuloksesta tulee todella tarkka. Koska permutaatiotes-

tissä saadaan selville kombinaatioiden arvojen otosvarianssi ja otoskeskiarvo, niin P-arvon luottamusvälejä ei päästä ratkaisemaan normaalijakaumalla, sillä normaalijakauma vaatii tiedon odotusarvosta ja varianssista. Jos käyttäjä on selvittänyt kombinaatioiden odotusarvon jollakin muulla tietokoneohjelmalla käymällä kaikki kombinaatiot tasan kerran läpi, niin siinä tapauksessa normaalijakaumaa voidaan kuitenkin hyödyntää.

P-arvolle saadaan todella tarkkoja arvioita muodostettua, vaikka aineisto ei noudattaisikaan mitään todennäköisyysjakaumaa. Se on samalla yksi MC -menetelmien vahvimmista puolista. Tässä työssä tutkittiin permutaatiotestiä keskiarvon vertailemisen kannalta. Aina kuitenkaan keskiarvo ei anna oikeanlaista kuvaa aineistosta, jolloin permutaatiotestin tuloksista ei voida tehdä kunnon johtopäätöksiä. Tälläisiä tapauksia voivat olla muun muassa aineistot, joissa on muutama todella suuri arvo tai aineiston alkioiden arvot painottuvat selkeästi ääripäihin aineiston suurimman ja pienimmän arvon erotuksen ollessa hyvin suuri. Näissäkin tapauksissa voidaan saada järkeviä tuloksia aikaan lajittelemalla alkiot luokkiin, jolloin eri luokat saavat eri arvon ja nyt permutaatiotestiä pystytään taas hyödyntämään aineiston tutkimisessa. Aineistoja itse muokatessa on kuitenkin hyvä muistaa pitää kriittinen näkökulma tuloksia arvioidessa, sillä muokatun aineiston tulokset eivät välttämättä kerro mitään järkevää alkuperäisestä aineistosta.

Vaikka käyttäjä ei tietäisi kombinaatioiden arvojen odotusarvoa, niin on tapoja arvioida P-arvon luottamusväliä tietyllä riskitasolla α. Käyttäjä voi laskea karkean luottamusvälin P-arvolle käyttämällä testistä saatua P-arvoa odotusarvona. Jos toistomäärä n on riittävän suuri, niin testin antama P-arvo, joka siis ilmaisee samalla kombinaatioiden arvojen odotusarvon, ei yleensä eroa ”oikeasta” odotusarvosta kovin suuresti. Käyttämällä testistä saatua P-arvoa odotusarvona ja laskemalla varianssi lauseesta 4, saadaan P-arvon luottamusväli laskettua normaalijakauman kautta. Arvio P-arvon luottamusvälistä voidaan tehdä myös hyödyntämällä keskeistä rajaarvolausetta. Tässä menetelmässä permutaatiotesti suoritetaan x kertaa, ja lasketaan testien antamien P-arvojen perusteella luottamusväli hyödyntämällä lausetta 6. On kuitenkin hyvä muistaa, että näillä menetelmillä ratkaistut luottamusvälit eivät ole päteviä ja niiden tarkoitus on ainoastaan muodostaa kuva testaajalle P-arvon käyttäytymistä. Tästä huolimatta näillä menetelmillä ratkaistut luottamusvälit näyttävät toimivan hyvin, sillä testasin mallinnusohjelmaa satoja kertoja tiedostoilla, joiden oikean odotusarvon tiesin, ja σ ja 2σ luottamusvälien sisään osui lähes oikean verran testituloksien P-arvoista. Tätä asiaa olisi hyvä tutkia lisää, koska jos P-arvoista saadaan muodostettua luotettavia luottamusvälejä, voidaan sitä tulevaisuudessa hyödyntää erityisesti tilastollisessa testaamisessa.

6. YHTEENVETO

Monte Carlo -menetelmät soveltuvat hyvin monenlaisten ongelmien ratkaisemiseen, koska MC -menetelmät ovat usein matemaattisesti yksinkertaisia sekä kevyitä suorittaa tietokoneohjelmilla. Menetelmien hyvänä puolena on myös se, että aineistojen ei tarvitse noudattaa mitään tunnetta todennäköisyysjakaumaa, jotta niitä voidaan käsitellä. Monissa menetelmissä toistomäärät saadaan kasvatettua menetelmien yksinkertaisuuden vuoksi suuriksi, jolloin saadut likiarvot ovat usein riittävän tarkkoja, jotta menetelmiä kannattaa hyödyntää.

Monte Carlo -permutaatiotestissä aineistosta valitaan vertailuryhmä, lasketaan vertailuryhmän keskiarvo ja koko ja päätetään toistomäärä. Nämä tiedot syötetään mallikoodiin, joka vertaa satunnaisesti arvottujen alkioiden keskiarvoa vertailuryhmän keskiarvoon ja testin lopussa ohjelma antaa käyttäjälle P-arvon, joka ilmaisee sen todennäköisyyden, että lopputulokseen päästäisiin olettamalla vertailuryhmän keskiarvo yhtä suureksi koko muun populaation keskiarvon kanssa. Kaiken kaikkiaan permutaatiotesti on toimii siis loistavana apuvälineenä aineistojen tutkimisessa, etenkin sellaisten aineistojen, jotka eivät noudata tunnettuja todennäköisyysjakaumia.

LÄHTEET

[1] K.-H.Jöckel, Bremen Institute: ”Finite sample properties and asymptotic efficiency of Monte Carlo tests,” The Annals of Statistics, vol. 14, no. 1, pp. 336–347, 1986

[2] L.S. Jun, Monte Carlo strategies in scientific computing, 1st ed. Spinger, 2008.

[3] H. Malvin, K. Kalos, P.A. Whitlock, Monte Carlo Methods, 2nd ed. WILEY-VCH Verlag GmbH & Co., 2008.

[4] M. Leino, ”Kvanttipisteiden Mallintaminen Polkuintegraali-Monte Carlo -menetelmällä,” Diplomityö, Tampereen teknillinen korkeakoulu, 2002

[5] A. Perttula, K. Vattulainen, T. Suurhasko, ”Todennäköisyyslaskenta -opintominoniste,” Tampereen teknillinen yliopisto, Versio 9/2012.

[6] J.Kangasaho, J.Mäkinen, J.Oikkonen, J.Paasonen, M.Salmela, J.Tahvanainen, ”Pitkä matematiikka 13: Differentiaali ja integraalilaskennan jatkokurssi,” Werner Söderström Osakeyhtiö, 1st ed., 2007

[7] F.C. Huang, Monte Carlo Integration -luentomoniste, UC Berkeley, 14.9.2009

[8] K. Ruihonen, Tilastomatematiikka -opintomoniste, 2011.

[9] J.Kangasaho, J.Mäkinen, J.Oikkonen, J.Paasonen, M.Salmela, J.Tahvanainen, ”Pitkä matematiikka 6: Todennäköisyys ja tilastot,” Werner Söderström Osakeyhtiö, 3rd ed., 2009.

[10] B.F.J. Manly, ”Randomization, Bootstrap and Monte Carlo Methods in Biology,” Taylor and Francis Group, 3rd ed., 2007.

LIITE 1. MONTE CARLO -PERMUTAATIOTESTIN MALLINNUSKOODI

```
function P_arvo = permutaatiotesti(K,L,N)
prompt = 'Syötä käsiteltävän tiedoston nimi!';
tiedosto = input(prompt,'s');
%Kysytään käyttäjältä tiedosto, josta otospopulaation tiedot luetaan
A = importdata(tiedosto);
% Luetaan tiedot syötteen tiedoston 1.sarakkeesta
B = A.';     % Muutetaan sarake vektoriksi --> helpompi käsitellä tietoa
L1 = length(B);      % permutaatiotesti laskee juuri tehdyn vektorin pituuden
E=0;
for i=1:N; % Luodaan silmukka, joka suorittaa N-kertaa L-alkioisen joukon arvonnan
    c = randperm(L1,L);     %Randperm varmistaa, että samaa alkiota ei valita useampaa kertaa samassa arvonnassa
    B2 = B(c);
    X = mean(B2); % Lasketaan keskiarvo saadulle arvontajoukolle

    if K<X  %Verrataan testin laskemaa keskiarvoa K:n arvoon ja jos ehto K < S toteutuu --> E:tä kasvatetaan yhdellä
        E=E+1;
    else
        E=E+0;
    end
end
P_arvo=E/N;  %Kun silmukka on käyty läpi (N-kertaa), niin lasketaan p-arvo (P) jakamalla E testin toistomäärällä N
J = num2str(P_arvo);
X=['P-arvo tarkastelujoukolle on ',J];
disp(X);      %Tulostetaan P-arvo näytölle
```