# Keskeisen raja-arvolauseen sovelluksia aikasarja-analyyseissä

Ari Koski

Pro gradu -tutkielma
Tammikuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Apulaisprofessori Joni Virta
Professori Henri Nyberg

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Ari Koski: Keskeisen raja-arvolauseen sovelluksia
aikasarja-analyyseissä
Pro gradu -tutkielma, 67 s., 16 liites.
Tilastotiede
Tammikuu 2026

---

Tässä tutkielmassa tarkastellaan keskeisen raja-arvolauseen soveltamista aikasarja-analyyseissä tilanteissa, joissa klassisen keskeisen raja-arvolauseen riippumattomuusoletus ei päde. Tarkastelu kohdistuu stationaarisiin AR-, MA- ja ARMA-malleihin, sekä tapauksiin joissa mallien parametrit ovat lähellä stationaarisuuden raja-arvoa. Menetelminä käytetään teoreettista analyysiä ja simulaatiokokeita, joissa arvioidaan asymptoottisiin tuloksiin perustuvien luottamusvälien peittotodennäköisyyksiä eri otoksilla ja parametrivalinnoilla.

Tulokset osoittavat, että keskeinen raja-arvolause pätee stationaarisissa aikasarjamalleissa myös riippuvuuden vallitessa. Tällöin luottamusvälien peittotodennäköisyydet lähestyvät nimellistasoa otoskoon kasvaessa. MA- ja ARMA-mallit tuottavat suhteellisen hyviä tuloksia jo pienillä otoksilla, edellyttäen että kääntyvyys- ja stationaarisuusehdot täyttyvät. Sen sijaan autoregressiivisissa malleissa tulokset ovat herkkiä parametrien arvoille stationaarisuuden raja-arvon läheisyydessä, mikä heikentää peittotodennäköisyyksiä pienillä otoksilla.

Avainsanat: keskeinen raja-arvolause, aikasarja-analyysi, ARMA-mallit, stationaarisuus, kääntyvyys.

# Sisällys

# 1 Johdanto

Tässä tutkielmassa tarkastellaan keskeisen raja-arvolauseen (KRL) soveltamista aikasarja-analyysiin. Tarkastelun kohteena ovat yleisesti käytetyt aikasarjamallit AR($p$), MA($q$) ja ARMA($p, q$), jotka alan kirjallisuudessa tyypillisesti esitetään perustapauksina ennen siirtymistä monimutkaisempiin malleihin. Näiden mallien avulla voidaan havainnollistaa KRL:n keskeiset edellytykset ja vaikutukset selkeästi, ilman että monimutkaisemmat rakenteet hämärtäisivät liikaa ilmiön olennaisia piirteitä. Tässä mielessä ne toimivat luontevana lähtökohtana analyysille.

Perinteinen keskeinen raja-arvolause edellyttää, että havaintomuuttujat ovat toisistaan riippumattomia ja samoin jakautuneita. Aikasarjoissa riippumattomuusoletus ei kuitenkaan päde, sillä havainnot ovat ajallisesti riippuvia. Tällöin havaintojen välinen yhteisvaihtelu vaikuttaa keskiarvon varianssiin ja siten asymptoottisten approksimaatioiden toimivuuteen. Työssä pyritään tarjoamaan intuitiivinen ja teoreettinen ymmärrys siitä, miten keskeinen raja-arvolause toimii erityisesti stationaarisissa aikasarjamalleissa.

Työ alkaa historiaosuudella, jossa käsitellään asymptoottisten menetelmien ja keskeisen raja-arvolauseen kehitykseen keskeisesti vaikuttaneita henkilöitä, sekä aikasarja-analyysin varhaisia kehitysvaiheita. Henkilövalinnat on rajattu työn laajuus huomioon ottaen matemaatikoihin ja tilastotieteilijöihin, jotka ovat olleet keskeisessä roolissa menetelmien kehityksessä.

Teoriaosuudessa tarkastellaan keskeisen raja-arvolauseen roolia aikasarjamalleissa ja havainnollistetaan sen soveltuvuutta stationaarisissa prosesseissa. Teoreettisia tuloksia täydennetään visuaalisilla esimerkeillä, joissa normaalijakauma toimii asymptoottisena approksimaationa. Visualisoinnit perustuvat erilaisiin parametrivalintoihin, mutta näissä odotusarvo ja varianssi pidetään vakioina arvoissa $\mu = 0$ ja $\sigma^2 = 1$.

Työn simulaatio-osiossa puolestaan teoriaosuudessa johdettuja tuloksia tarkastellaan yksityiskohtaisemmin 95 %:n luottamusvälien peittotodennäköisyyksien avulla useilla eri parametrikokoonpanoilla ja variansseilla. 95 %:n luottamusvälit muodostetaan hyödyntämällä mallien asymptoottisesta varianssista saatua keskihajontaa. Simulaatioiden tavoitteena on analysoida mallien käyttäytymistä otoskoon kasvaessa, eli tilanteessa jossa $n \to \infty$, sekä arvioida, missä määrin luottamusvälien peittotodennäköisyydet lähestyvät nimellistasoa 0.95. Simulaatioissa tarkastellaan myös tilanteita, joissa normaaliapproksimaatio ei ole riittävän tarkka, esimerkiksi pienen otoskoon vuoksi, liikuttaessa lähellä epästationaarisuuden raja-arvoa 1, tai kääntyvyysehdon rikkoutuessa.

Tutkielman laadinnassa on hyödynnetty ChatGPT-tekoälytyökalua (versiot GPT-3.5 ja GPT-4.0 ) tiedonhakuun ja alustavan tekstin hahmotteluun.

# 2 Merkinnöistä

Työssä yleisesti käytettyjä merkintöjä

- $\bar{X_n} = \frac{1}{n}\sum_{t=1}^{n} X_t$: Otoskeskiarvo
- $\mu = E(X_t)$: Stationaarisen prosessin odotusarvo
- $\sigma^2 = Var(X_t)$: Prosessin varianssi
- $\mathcal{N}(0, \sigma^2)$: Normaalijakauma, jonka odotusarvo 0 ja varianssi $\sigma^2$
- $n$: Otoskoko
- $\varepsilon_t$: Valkoisen kohinan termi/ virhetermi hetkellä $t$
- $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$: Normaalijakautunut virhetermi
- $X_t$ tarkasteltavan aikasarjaprosessin arvo hetkellä $t$
- $\phi_i$: AR($p$)-prosessin autoregressiivinen kerroin, jossa $p$ on AR-prosessin kertaluku
- $\theta_i$: MA($q$)-prosessin kerroin, jossa $q$ on MA-prosessin kertaluku
- $n \to \infty$: Otoskoko lähestyy ääretöntä
- $\xrightarrow{d}$: Suppeneminen jakaumamielessä (eng. In distribution)
- $\xrightarrow{p}$: Lähestyminen todennäköisyysmielessä (eng. Convergence in probability)
- $Y_0 \xrightarrow{p} 0, \quad \text{kun} \quad n \to \infty$: Satunnaismuuttuja lähestyy todennäköisyydessä kohti nollaa, kun otoskoko lähestyy ääretöntä
- $\gamma_k$: Autokovarianssi viiveellä $k$
- $\rho_k$: Autokorrelaatio viiveellä $k$
- $Var_{Asymp}$: Asymptoottinen varianssi
- i.i.d. Riippumaton ja samoin jakautunut satunnaismuuttuja (eng. Independent and identically distributed)
- $Y_t \sim$ i.i.d. $\mathcal{N}(0, 1)$, jossa $Y_t$:t ovat normaalijakautuneita i.i.d satunnaismuuttujia odotusarvolla $\mu = 0$ ja varianssilla $\sigma^2 = 1$.

# 3 Keskeisen raja-arvolauseen kehitykseen vaikuttaneita henkilöitä

Todennäköisyyslaskennan ja keskeisen raja-arvolauseen (KRL) kehityshistoriassa eräs keskeinen linja kulkee yksittäisten epävarmojen tapahtumien tarkastelusta kohti suurten joukkojen ennustettavaa käyttäytymistä. Varhaisimmat tutkijat, kuten Jacob Bernoulli ja Abraham de Moivre, keskittyivät usein yksittäisiin tilanteisiin, esimerkiksi peleihin, vakuutuslaskelmiin tai yksittäisiin kokeisiin, sekä niihin liittyvien todennäköisyyksien määrittämiseen. Tällaiset tilanteet olivat konkreettisia, rajattuja. Vähitellen kuitenkin huomattiin, että suurten lukumäärien kautta alkaa paljastua säännönmukaisuutta. Yksittäinen tapahtuma voi olla satunnainen ja arvaamaton, mutta tuhansien vastaavien tapahtumien kokonaisuus alkaa noudattaa yleisiä lainalaisuuksia. Tämä oivallus johti keskeisen raja-arvolauseen (KRL) (*eng. Central Limit Theorem*) muotoutumiseen ja tulokseen, josta tuli yksi todennäköisyyslaskennan ja tilastotieteen peruspilareista. Seuraavissa osioissa tarkastellaan asymptoottisten menetelmien ja KRL:n kehityskaarta varhaisten ajattelijoiden, kuten Bernoullin, de Moivren ja Laplacen myötä.

Keskeisen raja-arvolauseen historia on monivaiheinen ja siihen liittyy suuri joukko henkilöitä ja tapahtumia. Tässä työssä keskitytään kuitenkin niihin keskeisimpiin kehityslinjoihin ja henkilöihin, joiden kautta aihe voidaan esittää mahdollisimman ymmärrettävästi tämän työn rajojen puitteissa. Vaikka esimerkiksi Carl Friedrich Gaussin (1777–1855) rooli normaalijakauman kehittäjänä on merkittävä ja normaalijakauma liittyy olennaisesti KRL:ään, keskityn ensisijaisesti niihin matemaatikoihin, jotka tutkivat suoremmin asymptoottisia tuloksia ja vaikuttivat keskeisimmin lauseen syntyyn. Moni muukin merkittävä tutkija, kuten Augustin-Louis Cauchy (1789–1857), olisi ollut mahdollinen tarkastelun kohde, mutta valintani pyrkivät rakentamaan loogisesti etenevän ja sisällöllisesti yhtenäisen kokonaisuuden.

Osiossa 3.2 siirrytään tarkastelemaan aikasarja-analyysejä KRL:n näkökulmasta ja siinä keskitytään aikasarja-analyysien kehityksen alkuun. Tarkastelu keskittyy Karl Pearsonin ja Udny Yulen varhaisiin kontribuutioihin. Lisäksi käsitellään suomalaisen Leo Törnqvistin asymptoottista sovellusta hänen teoksessaan *Aikasarjojen analyysi ja ennustaminen* (1974), joka pohjautuu Herman Woldin esimerkkiin. Osiossa pyritään tuomaan esiin kuinka asymptoottiset menetelmät ja keskeinen raja-arvolause loivat tilastolliselle ajattelulle teoreettisen perustan, jonka varaan myös aikasarja-analyysien myöhempi kehitys osaltaan rakentui. Aikasarja-analyysien synty sai kuitenkin alkunsa ennen kaikkea soveltavista tarpeista, erityisesti talouden ja fysiikan aloilla. Historiaosuuden päättää osio 3.3, jossa esitellään suomalainen matemaatikko Jarl Waldemar Lindeberg, joka tunnetaan KRL:n merkittävänä tutkijana ja kehittäjänä.

## 3.1 Asymptoottisten menetelmien pioneerit 1700-1827 luvuilla

### 3.1.1 Jacob Bernoulli

Kuva 1: Jacob Bernoulli *[27]*.

Jacob Bernoulli (1654–1705) kuului kuuluisaan sveitsiläiseen Bernoullin matemaatikkosukuun, josta nousi useita aikansa merkittävimpiä tutkijoita. Hän loi perustan todennäköisyyslaskennan käsitteelliselle rakenteelle ja hän pyrki luomaan systemaattisen ja loogisesti johdonmukaisen teorian epävarmuudesta. Sellaisen, joka voisi tukea päätöksentekoa tilanteissa, joissa varmuutta ei ole tarjolla. Jacob Bernoulli oli tiedon etsijä maailmassa, jossa varmuus oli harvinaista. Hän jätti jälkeensä näkemyksen siitä, mitä tarkoittaa tehdä järkeviä päätöksiä epävarmoissa tilanteissa.

Bernoulli työskenteli vuosien ajan pääteoksensa *Ars Conjectandi* parissa. Hän ei kuitenkaan ehtinyt saattaa kirjaa valmiiksi ennen kuolemaansa, ja se julkaistiin postuumisti vuonna 1713 *[37]*. *Ars Conjectandi* oli uraauurtava monella tapaa. Se sisälsi niin peliteoreettisia analyysejä kuin syvällisen pohdinnan todennäköisyyden luonteesta. Kirjan lopusta löytyy Bernoullin tunnetuin perintö, niin sanottu suurten lukujen laki, joka osoitti, että toistettujen satunnaiskokeiden suhteelliset frekvenssit lähestyvät pitkällä aikavälillä vastaavia teoreettisia todennäköisyyksiä (*[17]*, s. 225). Hän esitti muodollisesti, että jos suoritetaan suuri määrä riippumattomia kokeita, joiden onnistumistodennäköisyys on $p$, ja $h_n$ on onnistumisten suhteellinen frekvenssi, niin

$$P\{|h_n - p| \leq \epsilon\} > 1 - \delta \quad \forall \epsilon, \delta > 0 \quad \text{kun} \quad n > n(p, \epsilon, \delta),$$

mikä tarkoittaa, että ero $|h_n - p|$ on pienempi kuin $\epsilon$ mielivaltaisen pienellä todennäköisyydellä, kunhan $n$ on tarpeeksi suuri. Tämä matemaattinen tulos tunnetaan suurten lukujen lakina ja se muodostaa keskeisen perusperiaatteen tilastollisessa arvioinnissa. Myöhemmin De Moivre puolestaan kehitti tätä ideaa ja esitti tarkempia laskelmia ja approksimaatioita, jotka paransivat Bernoullin alkuperäistä arviota. Hänen työnsä pohjautui Bernoullin lauseen vahvistamiseen ja tarkentamiseen erityisesti suurten $n$-arvojen osalta (*[14]*, s. 14).

Bernoulli ei kuitenkaan tyytynyt pelkkiin asymptoottisiin tuloksiin. Hän halusi tietää konkreettisesti, kuinka suuri otoskoko $n$ tarvitaan, jotta suhteellinen frekvenssi on halutun etäisyyden sisällä todennäköisyydellä $1 - \delta$. Vaikka hän ei esittänyt eksaktia kaavaa nykyisessä muodossa, hänen tavoitteensa oli selvästi sama, eli asettaa alaraja otoskoolle niin, että saavutetaan ennalta määrätty tarkkuus (*[17]*, s. 264). Tätä hän lähestyi esimerkein, kuten vertaamalla 3000 valkoisen ja 2000 mustan kiven suhteita otettuihin otoksiin ja arvioimalla, kuinka suurella määrällä havaintoja voidaan saavuttaa varmuus tuloksen luotettavuudesta (*[35]*, s. 13–14).

Bernoullin lauseen alkuperäinen muoto on luonteeltaan asymptoottinen ja perustuu logaritmisiin funktioihin, kuten Hald (*[17]*, s. 264) esittää seuraavasti

$$P_k > \frac{c(k)}{c(k)+1}, \text{ jossa } c(k) = c(r,s,k) \wedge c(s,r,k)$$

ja

$$ln\, c(r,s,k) = \frac{k(r+1)+s}{r+s+1}\, ln\frac{r+1}{r} - ln(s-1).$$

Koska $ln\, c(r,s,k)$ on lineaarinen funktio $k$:stä, Hald toteaa, että äärimmäisten poikkeamien todennäköisyys laskee vähintään eksponentiaalisesti, kun $k \to \infty$. Tämän pohjalta Hald esittää Bernoullin lauseen modernissa notaatiossa seuraavilla annetuilla arvoilla $p$, $\varepsilon > 0$ ja $c > 0$, josta seuraa

$$P_{m\epsilon} > \frac{c}{c+1} \quad \text{kun} \quad n \geq n(p,\varepsilon,c) \vee n(q,\varepsilon,c),$$

jossa

$$n(p,\varepsilon,c) \geq \frac{(1+\varepsilon)m - q}{\varepsilon(p+\varepsilon)}, \quad m \geq \frac{\ln\left[\frac{c(q-\varepsilon)}{\varepsilon}\right]}{\ln\left[\frac{p+\varepsilon}{p}\right]},$$

ja $n$, $m$ ovat pienimmät positiiviset kokonaisluvut, jotka täyttävät nämä epäyhtälöt. Tässä $P_{n\epsilon}$ tarkoittaa todennäköisyyttä, että onnistumisten osuus $k/n$ poikkeaa arvosta $p$ enintään $\epsilon$, jossa $k$ on onnistumisten lukumäärä $n$ toistossa.

Bernoulli kutsui lähestymistapaansa "arvailun taidoksi" (*lat. Ars conjectandi, eng. The Art of Conjecturing*) ja näki sen järkevänä keinona toimia tilanteissa, joissa täysi varmuus on saavuttamaton. Hän kirjoitti, että tavoitteena on löytää toimintatapa, joka on parempi ja varmempi silloin, kun täydellistä tietoa ei ole saatavilla (*[35]*, s. 5), (*[7]*, s. 213). Lisäksi Bernoulli tarkasteli todennäköisyyksiä, jotka nykyisin esitetään binomijakauman muodossa

$$P(k \text{ onnistumista}) = \binom{n}{k} p^k (1-p)^{n-k},$$

ja osoitti, kuinka onnistumisten jakauma kasaantuu keskiarvon ympärille. Hän esimerkiksi pohti moraalisen varmuuden käsitettä, eli kuinka suuri todennäköisyys on riittävä, jotta voimme luottaa tulokseen. Tätä hän tarkasteli yllä mainitun kiviesimerkin yhteydessä konkreettisesti laskemalla, kuinka epätodennäköisiä tietyt havainnot olisivat, jos perusoletus olisi tosi. Kirjassaan *Ars conjectandi* hän lähestyy asiaa seuraavasti

$$\frac{301}{200} \quad \& \quad \frac{299}{200} \quad \text{tai} \quad \frac{3001}{2000} \quad \& \quad \frac{2999}{2000} \quad \text{tai...}$$

(*[35]*, s. 13-14), (*[7]*, s. 226). Taustalla on oletus, että mitä suuremmista luvuista on kysymys, sitä vähemmän todennäköistä on, että kyseessä on sattuma. Tämä havainnollisti hänen keskeistä väitettään siitä, että minkä tahansa annetun todennäköisyyden suhteen on todennäköisempää, että usein toistetuilla kokeilla ja suurella otoskoolla saatu suhteellinen frekvenssi osuu ennalta asetettujen rajojen sisään, kuin niiden ulkopuolelle. Toisin sanoen, kun havaintojen määrä $n$ kasvaa, havaittu osuus lähestyy todellista suhdetta niin, että suhteellinen poikkeama käy yhä epätodennäköisemmäksi. Tämä liittyi suurten lukujen lain pohdintaan, joka sanoo, että mitä enemmän havaintoja on, sitä lähempänä suhteelliset todennäköisyydet ovat todellisia teoreettisia todennäköisyyksiä. Auki laskettuna äskeinen pohdinta antaa

$$\frac{301}{200} = 1.505 \quad \& \quad \frac{299}{200} = 1.495 \quad \text{tai} \quad \frac{3001}{2000} = 1.501 \quad \& \quad \frac{2999}{2000} = 1.4995.$$

Tämä osoittaa, että pienillä otoksilla tulos voi vaihdella huomattavasti, koska sattuman vaikutus on suurempi. Suuremmilla otoksilla tulokset alkavat lähestyä todennäköisyyksiä, jotka olisivat olleet teoreettisesti odotettavissa, ja ero on tällöin pienempi. Näin ollen suuret otoskoot tarjoavat luotettavampia arvioita. Kuitenkin voi huomata, että satunnaisvaihtelun vaikutus ei katoa täysin edes suurilla otoksilla. Esimerkiksi suhde

$$\frac{3001}{2000} = 1.501$$

on edelleen hieman etäällä siitä, mitä olisi odotettavissa tilanteessa, jossa populaatiossa on 3000 valkoista ja 2000 mustaa kiveä. Tässä tapauksessa populaation koko on

$$N = 3000 + 2000 = 5000,$$

jolloin todennäköisyydet saada valkoinen tai musta kivi satunnaisessa nostossa ovat

$$P(\text{valkoinen}) = \frac{3000}{5000} = 0.6, \quad P(\text{musta}) = \frac{2000}{5000} = 0.4.$$

Näiden todennäköisyyksien suhde on

$$\frac{P(\text{valkoinen})}{P(\text{musta})} = \frac{0.6}{0.4} = 1.5.$$

Tämä suhde edustaa populaation sisäistä rakennetta. Valkoisia kiviä on 1.5 kertaa enemmän kuin mustia. Kun populaatiosta otetaan satunnaisotanta, voidaan odottaa, että otoksessa havaittava suhde valkoisten ja mustien määrien välillä lähestyy tätä arvoa, mutta ei välttämättä saavuta sitä täsmälleen, ainakaan yksittäisessä otoksessa. Tässä näkyy satunnaisuuden luonne. Vaikka otoskoko olisi suuri, yksittäinen otos voi yhä poiketa teoreettisesta todennäköisyydestä. Kuitenkin suurilla otosmäärillä nämä poikkeamat ovat harvinaisempia ja pienempiä. Näin ollen ei voida odottaa tarkkaa vastaavuutta todellisen todennäköisyyden ja otoksessa havaitun osuuden välillä, mutta voidaan kuitenkin odottaa yhä parempaa likimääräisyyttä otoskoon kasvaessa. Juuri tämän kaltaista ilmiötä alettiin myöhemmin ymmärtää tarkemmin todennäköisyysteorian kehityksen myötä. Bernoullin analyysi kuvaa olennaisesti suurten lukujen lakia. Mitä enemmän havaintoja tehdään, sitä lähempänä otoksessa havaittu osuus on todellista todennäköisyyttä. Myöhemmin tutkijat, kuten de Moivre ja Laplace laajensivat

tätä ajatusta edelleen kehittämällä keskeisen raja-arvolauseen, joka ei ainoastaan kuvaa osuuksien lähestymistä odotusarvoon, vaan myös selittää millaista jakaumaa nämä keskiarvot noudattavat, kun $n \to \infty$. Normaalijakauman kehittymisen ja Gaussin[1] tulosten myötä voitiin tätä ilmiötä tutkia entistä paremmin. Näin Bernoullin varhaiset tutkimukset muodostivat perustan koko todennäköisyysteorian myöhemmälle kehitykselle.

Jacob Bernoulli tarkasteli myös jatkuvan kasvun ilmiötä ja siihen liittyvää raja-arvoa $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$, jonka hän esitti *Ars Conjectandi* -teoksessa. Vaikka hän ei käyttänyt nykyisin tunnettua merkintää $e$ eikä määrittänyt raja-arvon tarkkaa arvoa, hän osoitti, että raja-arvo on olemassa ja äärellinen (*[26]* s. 393). Kyseinen raja-arvo nousi myöhemmin keskeiseksi osaksi eksponenttifunktion teoriaa.

Jacob Bernoullin käsitys todennäköisyydestä ei perustunut pelkästään suhteellisiin frekvensseihin, vaan hän korosti myös rationaalisen päättelyn merkitystä epävarmuudessa. Tältä pohjalta Bernoullista tuli modernin tilastollisen päättelyn filosofinen esikuva. Hän ei tyytynyt siihen, että sattuma on väistämätöntä, vaan pyrki ymmärtämään sen säännönmukaisuuksia ja muotoilemaan ne matemaattisen teorian muotoon. Suurten lukujen laki, jonka Poisson[2] vakiinnutti vasta noin 120 vuotta myöhemmin, muodosti loogisen lähtökohdan keskeisen raja-arvolauseen myöhemmälle kehitykselle, kun tutkijat alkoivat järjestelmällisesti tutkimaan otoskeskiarvon käyttäytymistä ja pyrkivät yhä yleisempiin tuloksiin, sekä kehittämään menetelmiä, jotka laajensivat ja yleistivät Bernoullin alkuperäisiä tuloksia.

[1]Johann Carl Friedrich Gauss (1777-1855) oli saksalainen matemaatikko, tähtitieteilijä ja fyysikko, joka tunnetaan erityisesti Gaussin käyrän, eli normaalijakauman kehittäjänä.

[2]Simeon Denis Poisson (1781–1840) oli ranskalainen matemaatikko joka kehitti Poissoninjakauman.

### 3.1.2 Abraham de Moivre

Kuva 2: Abraham de Moivre *[36]*.

Abraham de Moivre (1667-1754) oli ranskalainen todennäköisyyslaskentaan erikoistunut matemaatikko, joka vaikutti 1700-luvun alussa. Hänen työnsä vaikuttivat perustavanlaatuisesti todennäköisyyslaskennan kehittymiseen tieteelliseksi menetelmäksi.

De Moivren tarina alkaa Ranskasta, mutta hänen matemaattinen uransa rakentui Englannissa, jonne hän pakeni hugenottina[3] uskonnollista vainoa. Tämä ulkopuolisuuden kokemus näkyy kenties rivien välistä myös hänen työssään, sillä järjestelmässä, jossa epävarmuus hallitsee, etsitään kaavaa ja säännönmukaisuutta (*[17]*, s. 397). Eräs hänen tunnetuimmista saavutuksistaan on teos *The Doctrine of Chances [12]* (Suom. Todennäköisyyden oppi), joka ilmestyi ensimmäisen kerran vuonna 1718. De Moivre pyrki tuomaan systematiikkaa ja menetelmällistä selkeyttä todennäköisyyslaskentaan, jotta se olisi kaikkien helposti käsiteltävissä. Hän siis pyrki luomaan käytännönläheisyyttä aiheeseen (*[17]*, s. 404). Yksi teoksen keskeisistä oivalluksista liittyy todennäköisyyden määrittelyyn suhteena suotuisten tapahtumien ja kaikkien mahdollisten tapahtumien välillä. De Moivre kirjoittaa

*"The Probability of an Event is greater or less, according to the number of Chances by which it may happen, compared with the whole number of Chances by which it may either happen or fail."* (*[12]*, s. 1).

Sivulla 2 hän pohtii klassista tapausta mahdollisuuksien määrästä, jolloin tapahtuma voi toteutua, tai mahdollisuuksien määrästä, jolloin se voi joko tapahtua tai epäonnistua. Tämän voi kiteyttää muodollisesti

$$P(A) = \frac{m}{n},$$

jossa $m$ on suotuisten ja $n$ kaikkien mahdollisten tulosten määrä. De Moivre kirjoittaa

[3]Hugenotit olivat ranskalaisia protestantteja, jotka joutuivat vainojen kohteeksi vuosina (1562 - 1764).

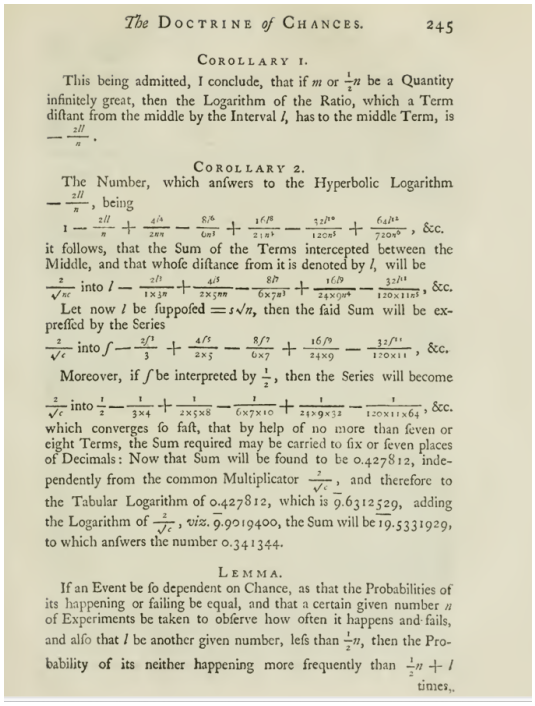
*"If upon the happening of an event, I be intitled to a sum of money, my expectation of obtaining that sum has a determinate value before the happening of the event. Thus, if I am to have* $10^{£}$ *in case of the happening of an event which has an equal probability of happening and failing, my expectation before the happening of the event is worth* $5^{£}$*: for I am precisely in the same circumstances as he who at an equal play ventures* $5^{£}$ *either to have 10, or to lose his 5. Now he who ventures* $5^{£}$ *at an equal play, is possessor of* $5^{£}$ *before the decision of the play."* (*[12]*, s. 2).

Tässä voiton odotusarvon ennen peliä voisi esittää muodossa $E(X) = p \cdot x + 0 \cdot (1-p) = p \cdot x = 0.5 \cdot 10 = 5$, jossa $p$ on tapahtuman todennäköisyys ja $x$ sen tuottama voitto. Tässä hän siis esitti odotusarvon käsitteen kyseiselle tapahtumalle.

Tämä viitekehys ei jäänyt vain pelipöydän ääreen. De Moivren ajattelu laajeni myös vakuutusmatematiikkaan ja elinkorkojen arviointiin (*[17]*, s. 508–515). De Moivre myös käsittelee sattuman ja suunnitelman erottamista toisistaan, erityisesti matemaattisen todennäköisyyden avulla. Hän esittää esimerkin korttipelistä, jossa arvioidaan, onko korttien järjestys sattumaa vai suunnitelman tulosta. Tämän pohjalta hän kehitti kaavoja, jotka mahdollistavat sattuman ja suunnitelman erottamisen laskennallisesti ja todennäköisyyksien perusteella (*[12]*, s. v–vi). De Moivre käytti Stirlingin[4] kaavan likimääräistä muotoa muodostaakseen laskentatavan binomitodennäköisyyksille suurilla otosmäärillä, jonka Stirling oli hänelle toimittanut yksityisesti. Tämä mahdollisti sen, että binomijakaumaa voitiin lähestyä eksponenttifunktion avulla tavalla, joka myöhemmin tunnistettiin normaalijakauman esimuodoksi (*[17]*, s. 470-492).

Erityisen havainnollinen esimerkki tästä löytyy *The Doctrine of Chances* -teoksen sivuilta 245-249, jonka sivut 245-246 esitetään kuvassa 3, de Moivre tarkastelee tapausta, jonka otoskoko on ($n = 3600$). Kokeen odotettujen onnistumisten määrä on ($n/2 = 1800$). Hän määrittelee tarkasteluvälin puolikkaan arvoksi ($\sqrt{n}/2 = 30$), jolloin odotetun arvon ympärille sijoittuva koko välin leveys on ($2 \cdot \sqrt{n}/2 = 60$). Hän tutkii todennäköisyyttä tapahtumalle, joka sijoittuu välille $[1770, 1830]$. Välit hän laskee käyttäen seuraavia kaavoja $\frac{1}{2} \cdot n - \frac{1}{2} \cdot \sqrt{n} = 1770$ ja $\frac{1}{2} \cdot n + \frac{1}{2} \cdot \sqrt{n} = 1830$. De Moivren mukaan tällaisen poikkeaman todennäköisyys on noin 0.682688. Kyseiseen lukemaan hän päätyi analysoimalla binomijakauman keskimmäisten termien logaritmisia etäisyyksiä (*[12] s. 245*), joista hän sai laskelmien perusteella summaksi 0.341344. Koska tämä on vain todennäköisyys yhdelle puolelle, saadaan koko välin todennäköisyydeksi $2 \cdot 0.341344 = 0,682688$. Tämän hän perusteli sillä, että tapahtuma, jolla on yhtä suuri todennäköisyys tapahtua tai jäädä tapahtumatta, ei ilmene useammin kuin $\frac{1}{2} \cdot n + \frac{1}{2} \cdot \sqrt{n}$ kertaa, eikä harvemmin kuin $\frac{1}{2} \cdot n - \frac{1}{2} \cdot \sqrt{n}$ kertaa. Tällöin tapahtuma esiintyy välillä $[1770, 1830]$ todennäköisyydellä $P(X \in \frac{1}{2} \cdot n \pm \frac{1}{2} \cdot \sqrt{n}) = 0.682688$ ja välin ulkopuoliset tulokset $1 - P(1770 \leq X \leq 1830) = 0.317312$ muodostavat loput todennäköisyydestä.

[4] James Stirling (1692-1770) oli skotlantilainen matemaatikko, joka tunnetaan Stirlingin approksimaatiosta.



*The* DOCTRINE *of* CHANCES. 245

COROLLARY 1.

This being admitted, I conclude, that if $m$ or $\frac{1}{2}n$ be a Quantity infinitely great, then the Logarithm of the Ratio, which a Term diſtant from the middle by the Interval $l$, has to the middle Term, is $-\frac{2ll}{n}$.

COROLLARY 2.

The Number, which anſwers to the Hyperbolic Logarithm $-\frac{2ll}{n}$, being

$1-\frac{ll}{n}+\frac{4l^4}{2nn}-\frac{8l^6}{6n^3}+\frac{16l^8}{24n^4}-\frac{32l^{10}}{120n^5}+\frac{64l^{12}}{720n^6}$, &c.

it follows, that the Sum of the Terms intercepted between the Middle, and that whoſe diſtance from it is denoted by $l$, will be

$\frac{2}{\sqrt{nc}}$ into $l-\frac{2l^3}{1\times 3n}+\frac{4l^5}{2\times 5nn}-\frac{8l^7}{6\times 7n^3}+\frac{16l^9}{24\times 9n^4}-\frac{32l^{11}}{120\times 11n^5}$, &c.

Let now $l$ be ſuppoſed $=s\sqrt{n}$, then the ſaid Sum will be expreſſed by the Series

$\frac{2}{\sqrt{c}}$ into $s-\frac{2s^3}{3}+\frac{4s^5}{2\times 5}-\frac{8s^7}{6\times 7}+\frac{16s^9}{24\times 9}-\frac{32s^{11}}{120\times 11}$, &c.

Moreover, if $s$ be interpreted by $\frac{1}{2}$, then the Series will become

$\frac{2}{\sqrt{c}}$ into $\frac{1}{2}-\frac{1}{3\times 4}+\frac{1}{2\times 5\times 8}-\frac{1}{6\times 7\times 10}+\frac{1}{24\times 9\times 32}-\frac{1}{120\times 11\times 64}$, &c.

which converges ſo faſt, that by help of no more than ſeven or eight Terms, the Sum required may be carried to ſix or ſeven places of Decimals: Now that Sum will be found to be 0.427812, independently from the common Multiplicator $\frac{2}{\sqrt{c}}$, and therefore to the Tabular Logarithm of 0.427812, which is $\overline{9}.6312529$, adding the Logarithm of $\frac{2}{\sqrt{c}}$, *viz.* $\overline{9}.9019400$, the Sum will be $\overline{19}.5331929$, to which anſwers the number 0.341344.

LEMMA.

If an Event be ſo dependent on Chance, as that the Probabilities of its happening or failing be equal, and that a certain given number $n$ of Experiments be taken to obſerve how often it happens and fails, and alſo that $l$ be another given number, leſs than $\frac{1}{2}n$, then the Probability of its neither happening more frequently than $\frac{1}{2}n+l$ times,

246 *The* DOCTRINE *of* CHANCES.

times, nor more rarely than $\frac{1}{2}n-l$ times, may be found as follows.

Let L and L be two Terms equally diſtant on both ſides of the middle Term of the Binomial $\overline{1+1}^n$ expanded, by an Interval equal to $l$; let alſo $f$ be the Sum of the Terms included between L and L together with the Extreams, then the Probability required will be rightly expreſſed by the Fraction $\frac{f}{2^n}$; which being founded on the common Principles of the Doctrine of Chances, requires no Demonſtration in this place.

COROLLARY 3.

And therefore, if it was poſſible to take an infinite number of Experiments, the Probability that an Event which has an equal number of Chances to happen or fail, ſhall neither appear more frequently than $\frac{1}{2}n+\frac{1}{2}\sqrt{n}$ times, nor more rarely than $\frac{1}{2}n-\frac{1}{2}\sqrt{n}$ times, will be expreſſed by the double Sum of the number exhibited in the ſecond Corollary, that is, by 0.682688, and conſequently the Probability of the contrary, which is that of happening more frequently or more rarely than in the proportion above aſſigned will be 0.317312, thoſe two Probabilities together compleating Unity, which is the meaſure of Certainty: Now the Ratio of thoſe Probabilities is in ſmall Terms 28 to 13 very near.

COROLLARY 4.

But altho' the taking an infinite number of Experiments be not practicable, yet the preceding Concluſions may very well be applied to finite numbers, provided they be great: for Inſtance, if 3600 Experiments be taken, make $n=3600$, hence $\frac{1}{2}n$ will be $=1800$, and $\frac{1}{2}\sqrt{n}=30$, then the Probability of the Event's neither appearing oftner than 1830 times, nor more rarely than 1770, will be 0.682688.

Kuva 3: Abraham de Moivren laskelma kirjasta *The Doctrine of Chances*, sivut 245–246.

Tämä on vaikuttava osoitus siitä, kuinka de Moivre pystyi antamaan likimääräisiä mutta tarkkoja arvioita todennäköisyyksistä suurilla otosmäärillä. Vaikka hän ei vielä käyttänyt normaalijakauman käsitettä muodollisesti, eikä hänellä ollut standardinormaalijakauman taulukoita käytettävissä, niin hänen menetelmissään näkyy jo intuitiivinen ymmärrys siitä, miten suuret otoskoot johtavat tulosten keskittymiseen odotusarvon ympärille. Tämä on juuri se ilmiö, joka myöhemmin muodollistettiin keskeisessä rajaarvolauseessa.

### 3.1.3 Klassisen KRL:n soveltaminen de Moivren esimerkkiin

Myöhemmin osiossa 4 esitetään tarkemmin KRL:n perusmuoto. Tässä vaiheessa voidaan kuitenkin havainnollistaa, kuinka se soveltuisi de Moivren esimerkkiin. KRL:n mukaan, toisistaan riippumattomien ja samoin jakautuneiden satunnaismuuttujien otoskeskiarvot alkavat lähestyä normaalijakaumaa otoskoon kasvaessa. Lause voidaan esittää muodossa

$$\sqrt{n}(\bar{X}-\mu)/\sigma \xrightarrow{d} N(0,1), \quad \text{kun} \quad n\to\infty.$$

Tarkastellaan yllä käsiteltyä tilannetta, jossa de Moivre suoritti $n=3600$ Bernoullikoetta onnistumistodennäköisyydellä $p=0.5$. Esimerkissä käytettävät parametrit ovat odotusarvo $\mu=n\cdot p=1800$ ja keskihajonta $\sigma=\sqrt{n\cdot p(1-p)}=30$. De Moivre tarkasteli väliä

$$[1800-30,\ 1800+30]=[1770,\ 1830],$$

eli poikkeamaa $\pm\frac{1}{2}\cdot\sqrt{n}$. Tämä vastaa suhteellista väliä

$$\bar{X} \in \left[\frac{1770}{3600},\ \frac{1830}{3600}\right].$$

Lasketaan standardoitu muuttuja, joka käyttää keskeisen raja-arvolauseen perusperiaatteita, eli se näyttää, kuinka otoskeskiarvon jakauma lähestyy normaalijakaumaa, kun otoskoko kasvaa. Tässä on otettu huomioon myös otoskoko $n$ ja populaation keskihajonta $\sigma$, jotta voidaan standardoida otoskeskiarvo ja verrata sitä standardinormaalijakaumaan.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 0.5}{0.5/60} = 120(\bar{X} - 0.5).$$

Koska aiemmin tarkasteltu väli $[1770, 1830]$ vastaa suhteellisia osuuksia $\left[\frac{1770}{3600},\ \frac{1830}{3600}\right]$, sijoitetaan nämä luvut standardointikaavaan, jolloin välin päätepisteiksi saadaan

$$120\left(\frac{1770}{3600} - 0.5\right) = -1 \quad \text{ja} \quad 120\left(\frac{1830}{3600} - 0.5\right) = 1.$$

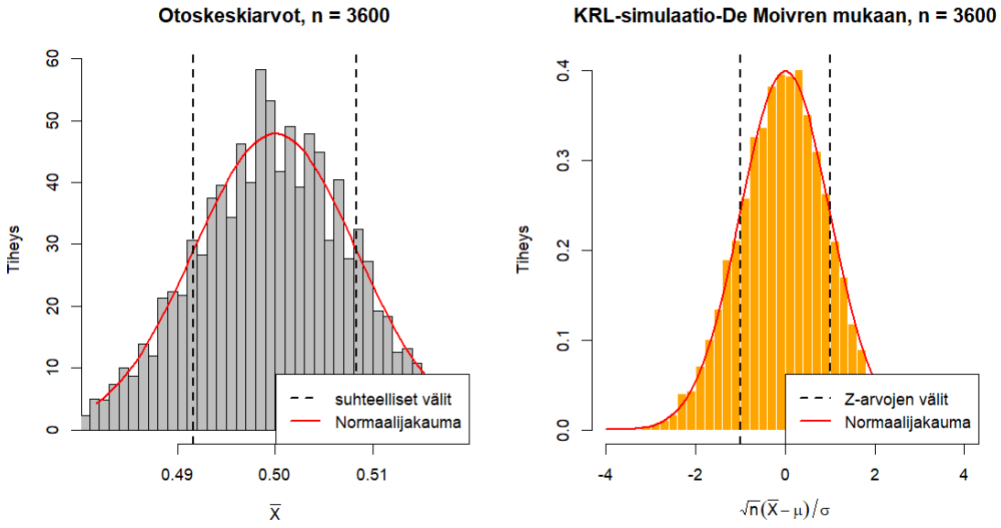
Tällöin normaalijakauman kertymäfunktion avulla puolestaan saadaan tulokseksi

$$P(-1 \leq Z \leq 1) = \Phi(1.0) - \Phi(-1.0) = 0.8413447 - 0.1586553 = 0,682689,$$

joka vastaa lähes täsmälleen de Moivren mainitsemaa arviota 0.682688. Tämä osoittaa, kuinka hänen intuitiivisesti määrittelemänsä väli ennakoi tarkasti normaalijakauman symmetriaa. Vaikka de Moivre ei tuntenut normaalijakaumaa muodollisesti, hänen analyysinsä kuvasi sen ominaisuuksia poikkeuksellisen tarkasti jo 1700-luvulla.

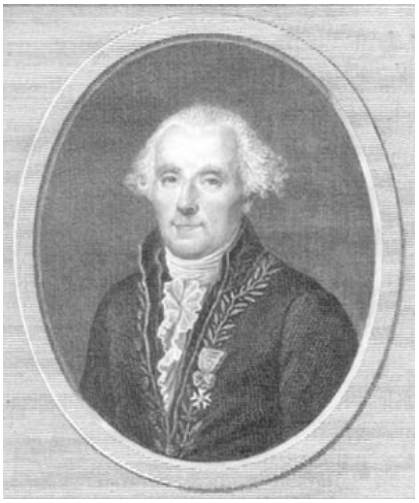
Lasketun todennäköisyyden havainnollistamiseksi tehtiin lisäksi simulaatio, joka esitetään kuvassa 4. Tässä toistettiin 10 000 kertaa 3600 Bernoulli-koetta ja laskettiin, kuinka moni otoskeskiarvo $\bar{X}$ osui välille $\left[\frac{1770}{3600},\ \frac{1830}{3600}\right]$. Simuloitu osuus oli $0,682689$, eli täsmälleen sama kuin saatu teoreettinen arvo. Tämä tukee hyvin KRL:n soveltuvuutta de Moivren tilanteeseen ja osoittaa, kuinka käytännön simulaatiot ja teoreettinen laskenta tukevat toisiaan.

Kuvassa 4 harmaan histogrammin päälle piirretyt katkoviivat näyttävät suhteelliset välit otoskeskiarvon jakaumasta. Tämän välin rajat on laskettu suhteellisesti onnistumismäärille ja harmaa kuvaaja havainnollistaa, kuinka otoskeskiarvot jakaantuvat suhteellisiin osuuksiin $\left[\frac{1770}{3600},\ \frac{1830}{3600}\right]$. Oranssi kuvaaja näyttää vastaavat Z-arvot, jotka on laskettu käyttäen KRL:n kaavaa $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Siinä standardoidut $Z$-arvot jakautuvat normaalisti $N(0,1)$, ja väli $Z = -1$ ja $Z = 1$ jakaa $Z$-arvojen jakauman.




Kuva 4: De Moivren KRL-esimerkki

### 3.1.4 Pierre-Simon Laplace



Kuva 5: Pierre-Simon Laplace *[18]*.

Pierre-Simon Laplace (1749-1827) oli ranskalainen matemaatikko, joka kunnostautui erityisesti matematiikan, tähtitieteen ja todennäköisyyslaskennan aloilla. Hän jatkoi Bernoullin ja De Moivren viitoittamaa tietä ja kehitti edelleen teoreettisia ja asymptoottisia menetelmiä.

Laplace syntyi Ranskassa 1750 ja vietti elämänsä ensimmäiset kuusitoista vuotta Normandiassa. Hänen isänsä halusi hänestä pappia, ja hän aloitti opiskelut Caenin yliopistossa 1766 taideaineiden tiedekunnassa tarkoituksenaan suuntautua teologian opintoihin myöhemmin. Kuitenkin kahden vuoden opintojen jälkeen hän suuntasi matematiikan opintoihin Pariisiin, jonne sai mukaansa suosituskirjeen Caenin yliopiston matematiikan opettajalta. Hieman myöhemmin hän aloitti matematiikan opettajan työt Ecole militaire -koulussa. Vuosien 1770- ja 1774 välillä Laplace teki suuren määrän merkittäviä kirjoituksia tähtitieteen, matematiikan ja todennäköisyyslaskennan saralla, ja vain 24-vuotiaana hänet valittiin Pariisin tiedeakatemian apujäseneksi ja myöhemmin täysjäseneksi vuonna 1785. Hän toimi myös politiikassa Napoleon Bonaparten nimittäessä hänet kuudeksi viikoksi sisäministeriksi ja myöhemmin kansleriksi. Ranskan restauraation jälkeen kuningas Ludvig XVIII teki hänestä Ranskan paronin ja myönsi hänelle markiisin arvonimen. Laplace asui koko loppuelämänsä Pariisissa *[18]*.

Laplace julkaisi merkittäviä tähtitieteeseen suuntautuvia teoksia, minkä jälkeen hän vuonna 1805 palasi jälleen todennäköisyyslaskennan ja tilastotieteen pariin, ja kirjoitti ehkä merkittävimmän siihen mennessä tehdyn todennäköisyyslaskentaan ja asymptoottisiin menetelmiin suuntautuvan teoksen *Theorie Analytique des Probabilites [22]*, jonka viimeinen painos lisäyksineen painettiin vuonna 1825. Hän teki kirjasta myös kansantajuisen version tieteellisen version rinnalle *[18]*. Kirja oli edistyksellinen monella tapaa ja siinä tutkittiin muun muassa otoksien käyttäytymistä ja muotoa, kun niiden koko kasvaa rajatta. Tämä lähestymistapa muodosti perustan sille, mitä myöhemmin kutsuttiin keskeiseksi raja-arvolauseeksi. Teos jakautuu kahteen eri kirjaosioon, joiden englanninkieliset käännökset on tehnyt Richard J. Pulskamp (*[23]. Kirja 1*) ja (*[24]. Kirja 2*). Erityisesti kirjassa 2 luvussa 3 "On the laws of probability, which result from the indefinite multiplication of events" käsitellään todennäköisyyden lakien ilmenemistä tapahtumien toistuvan kertauksen kautta. Esimerkiksi sivulla 285 Laplace esittää seuraavan kaavan, jota hän kutsuu kaavaksi $(o)$,

$$\frac{x+l}{n} - p = \frac{1+z}{n} = \frac{t\sqrt{2xx'}}{n\sqrt{n}} + \frac{z}{n}. \tag{1}$$

Kaava kuvaa tilannetta, jossa tutkitaan kuinka paljon tapahtuman $a$ esiintymiskertojen suhteellinen osuus voi poiketa sen todennäköisyydestä $p$, jolloin se antaa rajat joiden sisään havaittu suhteellinen osuus todennäköisesti jää,

$$\pm\frac{t\sqrt{2xx'}}{n\sqrt{n}} + \frac{z}{n},$$

jossa

$$\sqrt{2xx'} = n\sqrt{2p(1-p) + \frac{2z}{n}(1-2p) - \frac{2z^2}{n^2}}.$$

Rajojen väli on suuruusluokkaa $\frac{1}{\sqrt{n}}$, ja se pienenee $n$:n kasvaessa, eli kun $n \to \infty$.

Tämä tulos on pitkän laskelman lopputulos Laplacen teoksesta (*[22] s. 280-285*), jossa hän kehittää binomijakaumaan perustuvaa approksimaatiota. Laplace huomauttaa, että suurilla otoksilla tapahtumien suhteellinen esiintyvyys pysyy annettujen rajojen sisällä. Tämä oli merkittävä havainto, sillä normaalijakauman käsite ei tuolloin ollut vielä vakiintunut, mutta Laplacen laskelmat ennakoivat keskeisen raja-arvolauseen ideaa, jossa binomijakauma lähestyy suurilla otoksilla normaalijakaumaa.

Esimerkkinä hän antaa tapauksen, jossa poikien ja tyttöjen syntyvyyksien suhde on pojille 18 ja tytöille 17. Vuodessa on syntynyt yhteensä $n = 14000$ lasta. Hän haluaa arvioida todennäköisyyttä sille, että poikien syntymien määrä jää välille $[7037 - 7363]$. Tässä poikien todennäköisyys on $p = \frac{18}{17+18} = \frac{18}{35}$ ja tyttöjen $1 - p = 1 - \frac{18}{17+18} = \frac{17}{35}$. Poikien odotettu määrä $x = p \cdot n = \frac{18}{35} \cdot 14000 = 7200$ ja tyttöjen odotettu määrä $x' = (1 - p) \cdot n = (\frac{17}{35}) \cdot 14000 = 6800$. Tässä hän antaa poikkeamaksi luvun $l = 163$, joka saadaan laskemalla $l = \frac{x+l}{n} - p \cdot n = 163$. Seuraavassa kaavassa myös $z = 163$ ja $t = \frac{1}{\sqrt{n}} = 0.008451543$. Sijoittamalla nämä arvot kaavaan (1) saadaan poikien havaitun ja teoreettisen osuuden eron välille

$$\frac{7200 + 163}{14000} - \frac{18}{35} = \frac{1 + 163}{14000} = \frac{0.008451543 \cdot \sqrt{2 \cdot 7200 \cdot 6800}}{14000 \cdot \sqrt{14000}} + \frac{163}{14000} = 0.0116933.$$

Saatu arvo kuvaa poikien osuuden suhteellista poikkeamaa teoreettisesta odotusarvosta. Laplace laskee tämän todennäköisyyden olevan 0.994303, mikä R-ohjelmistolla tarkistettuna antaa arvon

```
> n <- 14000
> p <- 18 / 35
> mu <- p * n
> t <- 163
> sigma <- sqrt(n * p * (1-p))
> # Välin rajat
> a <- mu - t
> b <- mu + t
> # Todennäköisyys, että jää välien sisään
> prob <- pnorm(b, mean = mu, sd = sigma) - pnorm(a, mean = mu, sd = sigma)
> prob
[1] 0.9941546
```

Tämä vastaa hyvin Laplacen kirjassa laskemaansa arvoa. Approksimatiivinen todennäköisyys sille, että poikien syntymän määrä jää välille $[7037-7363]$ on siis $p = 0.9941546$.

Laplacen teos *Theorie Analytique des Probabilites* oli monella tapaa edistyksellinen ja jatkoi Bernoullin ja De Moivren viitoittamalla tiellä. Vaikka Bernoulli ja De Moivre olivat jo aikaisemmin käsitelleet vastaavanlaisia tapauksia, niin Laplace vei kehitelmät pidemmälle ja hänen laajassa työssään käsitellään tarkemmin ja analyyttisemmin suurten otoksien asymptoottista käyttäytymistä. Esimerkiksi ensimmäisen kirjaosion luvussa kolme "Theory of the approximations of formulas which are functions of large numbers" (*[22] s.128*), Laplace tutkii hyvin tarkasti parametrien asymptoottista käyttäytymistä äärettömyydessä.

## 3.2 Aikasarja-analyysien historiaa keskeisen raja-arvolauseen valossa

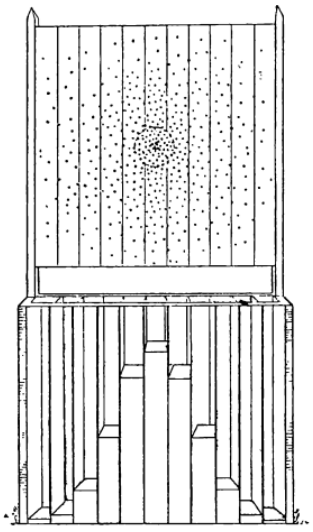
Aikasarja-analyysien historia on hyvin tuore ja se on kehittynyt vähitellen eri tutkijoiden töiden tuloksena viimeisen sadan vuoden aikana. Vaikka keskeinen raja-arvolause on tärkeä ja välttämätön osa aikasarja-analyysejä, niin sen vaikutusta aikasarja-analyyseihin

on historian saatossa tutkittu hyvin vähän, ja siitä on saatavilla niukasti lähdemateriaalia. Tämän vuoksi kattavan historiallisen yhteenvedon luominen juuri KRL:n valossa oli haastavaa. Kuten aikaisemmissa luvuissa olen käsitellyt, niin Bernoullin, De Moivren ja Laplacen kaltaisilla matemaatikoilla oli ratkaiseva osa aikasarja-analyysien ja keskeisen raja-arvolauseen kehittymiseen. Varsinaiset aikasarja-analyysit, kuten ne tänäpäivänä tunnemme, alkoivat hahmottua kuitenkin vasta 1900-luvun alussa.

Aikasarja-analyysien kehitykseen on vaikuttanut suuri joukko tutkijoita. Tämä osio on rajattu aikasarja-analyysien alkulähteisiin ja siinä käsitellään Karl Pearsonin[5] ja Udny Yulen panosta alan kehitykseen. Osion päättää Leo Törnqvist, joka oli suomalainen aikasarja-analyysien kehittäjä.

### 3.2.1 Aikasarja-analyysien alkusysäyksiä

Kun standardinormaalijakauma oli Carl Friedrich Gaussin (1777–1855) tutkimusten myötä vakiinnuttanut asemansa, sen käyttö alkoi vähitellen muodostua yleiseksi käytännöksi. Monet myöhemmät tutkimukset keskittyivät normaalijakaumaan ja sen olemassaolon empiiriseen todentamiseen. Esimerkiksi Karl Pearson ja Aleksandr Lyapunov[6] tarkastelivat ampumatuloksia ja havaitsivat, että luodinreikien jakauma seurasi normaalijakauman muotoa. Lyapunovin tutkimusryhmä käytti tässä tykinkuulien osumapisteiden hajontaa, joilla he havainnollistivat KRL:n merkitystä stokastisille prosesseille *[21]*.



Kuva 6: Pearsonin ampumatesti vuodelta 1897 *[21]*.

[5]Karl Pearson (1857–1936) oli englantilainen tilastotieteilijä ja matemaattisen tilastotieteen uranuurtaja. Hän perusti maailman ensimmäisen yliopistollisen tilastotieteen laitoksen University College Londoniin vuonna 1911.

[6]Aleksandr Mikhailovich Lyapunov (1857–1918) oli venäläinen matemaatikko, joka kehitti keskeistä raja-arvolausetta (KRL) niin, ettei satunnaismuuttujien $X_i$ tarvinnut olla identtisesti jakautuneita.

Kuva 6 esittää Karl Pearsonin ampumatestiä vuodelta 1897. Testissä luodinreikien sijaintien hajonta keskikohdasta muodostaa normaalijakauman, kun ne pudotetaan alla olevaan säiliöön. Pearson kehitti myöhemmin tunnetun $\chi^2$-testin (*Khiin-neliötesti. Eng. Chi-Square test*) vuonna 1900, mutta on todennäköistä, että hän oli aloittanut sen kehittelyn jo aiemmin. Ampumatestissä käytetäänkin todennäköisesti juuri tämän tyyppistä menetelmää. Tutkimusten juuret ulottuvat kuitenkin paljon pidemmälle, sillä jo vuonna 1820 Laplace pyrki perustelemaan pienimmän neliösumman menetelmän kahdelle tuntemattomalle muuttujalle suurten havaintomäärien tapauksessa, käyttäen minimivirheen periaatetta *[31]*.

Pearsonia pidetään matemaattisen tilastotieteen perustajana. Hänen tutkimuksensa korrelaatiokertoimien, varianssin ja hajonnan sekä $\chi^2$-testin parissa loivat pohjan nykyaikaisille aikasarja-analyysien menetelmille. Vaikka Pearson ei itse varsinaisesti kehittänyt aikasarja-analyysejä, niin hänen kehitelmänsä mahdollistivat, että myöhemmät tutkijat pystyivät kehittämään varsinaiset aikasarja-analyysin teoriat. Pearsonin kehittämät menetelmät pohjautuvat vahvasti keskeiseen raja-arvolauseeseen. Esimerkiksi KRL selittää, miksi $\chi^2$-testi toimii suurilla otoksilla, sillä summan neliöt normalisoituvat ja lähestyvät normaalijakaumaa. Vastaavasti Pearsonin korrelaatio- ja varianssilaskelmissa KRL mahdollistaa hypoteesitestien tekemisen, sillä keskiarvojen ja summien jakaumat aikasarjoissa lähestyvät normaalijakaumaa, kun otoskoko kasvaa. Tämä on keskeistä myös aikasarja-analyyseissä. Vaikka havaintopisteet voivat olla osittain riippuvaisia ajassa, niin keskiarvojen ja summien jakaumat voivat lähestyä normaalijakaumaa, kun aikasarja on stationaarinen ja riittävän pitkä. Näin Pearsonin menetelmät, kuten korrelaatio- ja varianssilaskelmat voidaan soveltaa autokorrelaation arviointiin, trendien tunnistamiseen ja testien tekemiseen aikasarjoissa, mikä muodostaa pohjan monille nykyisille aikasarja-malleille.

Seuraavassa ote Pearsonin kirjeestä hänen ystävälleen W.H Macaulaylle[7] vuodelta 1895. Kirje liittyi vinoutuneiden jakaumien tutkimiseen ja siihen sopivan järjestelmän löytämiseen.

*"There is a long tale as to the skew curves. Edgeworth came to me with some skew price curves nearly 18 months or two years ago and asked me if I could discover any means of dealing with skewness. I had come to skewness also in my Gresham lectures. I went to him in about a fortnight and said I think I have got a solution out, here is the equation, and told him my chief (assumed) discoveries. I further said I don't intend to publish till I have illustrated every point from practical statistics."[30].*

### 3.2.2 Udny Yule ja autoregressiivisen mallin kehittäminen

Eräs merkittävimmistä aikasarja-analyyseihin vaikuttaneista tilastotieteilijöistä on Udny Yule (1871-1951). Yule valmistui University College Londonista (UCL) insinööriksi. Vuonna 1893 Carl Pearson nimitti hänet UCL:ään assistentikseen. Yulen ensimmäinen tilastotiedettä käsittelevä artikkeli ilmestyi vuonna 1895 ja vuonna 1911 hän julkaisi teoksen *Introduction to the Theory of Statistics [44]*, joka oli yksi keskeisimpiä matemaattista tilastotiedettä käsitteleviä teoksia seuraavien 40 vuoden ajan *[40]*. Ehkä kuitenkin hänen uransa tunnetuimpia saavutuksia aikasarja-analyysien osalta oli autoregressiivisen mallin kehittäminen vuonna 1927, joka mullisti tilastotieteen alan. Hän

[7]William Herrick Macaulay (1853-1936) oli englantilainen matemaatikko.

julkaisi muun muassa auringonpilkkusarjoja tutkivassa artikkelissaan *On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers [43]*, toisen kertaluvun autoregressiivisen AR(2)-mallin, kun hän analysoi auringonpilkkusarjoja. AR($p$) -nimitys tuli myöhemmin. Hänen työssään ei varsinaisesti kerrota KRL:n merkityksestä, mutta se on silti läsnä taustalla (Myöhemmin tässä työssä käydään läpi kuinka keskeinen raja-arvolause on välttämätön ja oleellinen osa AR-malleja).

Esimerkiksi artikkelin alussa Yule esittää harmonisen liikkeen päällekkäisiä satunnaisia virheitä. Esimerkkinä hän käyttää nopanheitolla saatuja satunnaisia virheitä. Hän pyrkii tuomaan esiin, kuinka todelliset häiriöt voivat muuttaa amplitudia ja vaihetta, mikä tekee niiden analysoinnista vaikeampaa. Nämä virheet ovat riippumattomia ja odotusarvoltaan nollan ympärillä, ja niiden summat käyttäytyvät KRL:n mukaan. Eli pitkän sarjan summien jakauma lähestyy normaalijakaumaa riippumatta yksittäisten virheiden alkuperäisestä jakaumasta. Tämän vuoksi suuretkin satunnaiset virheet keskimäärin tasoittuvat, ja periodogrammi pystyy silti havaitsemaan alkuperäisen harmonisen jakson, kun havaintoja on riittävästi, kuten kuvassa 7. Sen sijaan, kuten kuvassa 8 esitetään, todelliset häiriöt eivät ole yksinkertaisia satunnaisia poikkeamia, vaan vaikuttavat jatkuvasti amplitudiin ja vaiheeseen. Tällöin KRL ei päde samalla tavalla, sillä KRL:n toimivuus perustuu riippumattomien satunnaismuuttujien summien käyttäytymiseen, kun taas jatkuvat häiriöt eivät ole yksittäisiä virheitä, vaan systemaattisia muutoksia.

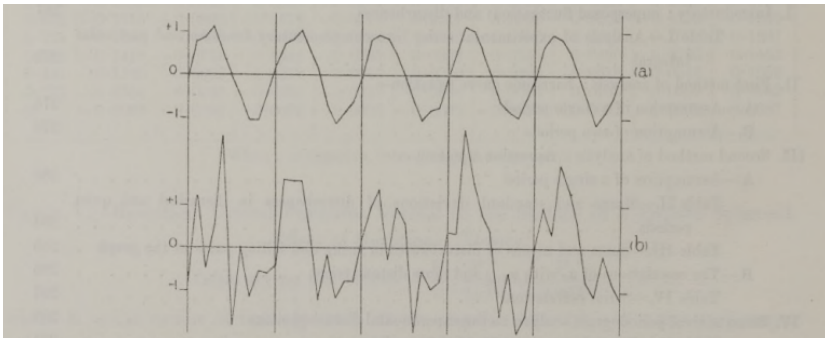
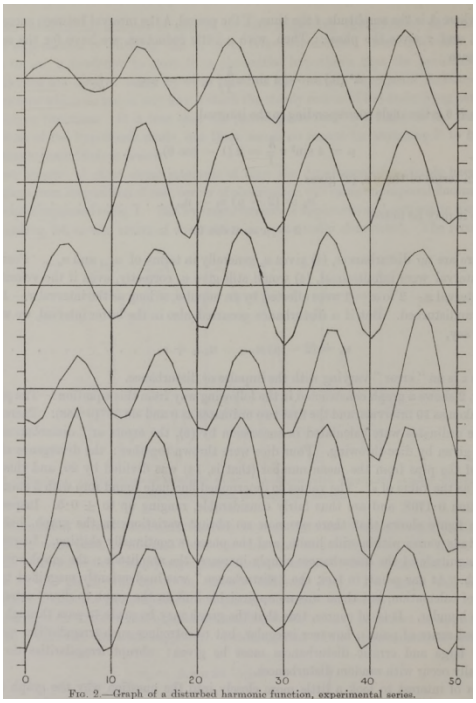


FIG. 1.—Graphs of simple harmonic functions of unit amplitude with superposed random fluctuations: (*a*) smaller fluctuations, (*b*) larger fluctuations.

Kuva 7: Yulen harmoninen liike, jossa satunnaiset virheet tasoittuvat otoskoon kasvaessa. *[43]*.




Kuva 8: Yulen harmoninen liike, jossa todelliset häiriöt vaikuttavat jatkuvasti amplitudiin ja vaiheeseen. *[43]*.

Yulen työ autoregressiivisten mallien parissa mullisti alan. Tätä seurasi Eugen Slutskyn[8] kanssa tehdyt tutkimukset satunnaishäiriöiden lineaarisista muunnoksista, sekä Yulen ja Oskar Andersonin[9] aikasarjaluokittelu. Myös Andrei Kolmogorovin[10] todennäköisyysteorialla oli suuri vaikutus aikasarja-analyysien kehityksessä. Tämän jälkeen kehitystä jatkoivat Aleksandr Khinchinin[11] ja Herman Woldin[12] työt stationaaristen ja stokastisten prosessien parissa *[21]*. Myöhemmin muun muassa Herman Wold ja Andrei Kolmogorov kehittivät stokastisten prosessien teoriaa, vahvistaen KRL:n keskeistä roolia aikasarja-analyysissä.

1900-luvun alkupuoliskolla aikasarja-analyysien tutkimuksessa ei vielä keskitytty niinkään asymptoottisiin menetelmiin, vaikka asymptoottisia menetelmiä oli tutkittu jo noin 200 vuotta. Vaikka varhaisimmat tilastolliset tutkimukset eivät vielä suoraan kohdistuneet aikasarjoihin, ne loivat pohjan normaalijakauman ja summien käyttäytymisen ymmärtämiselle. Kuten aiemmin tässä osiossa kerrottiin, 1800-luvun lopulla Karl Pearson sovelsi normaalijakaumaa käytännön mittauksiin, kuten ampumatesteihin, ja kehitti $\chi^2$-testin. KRL selittää, miksi summat ja keskiarvot käyttäytyvät normaalisti suurilla otoksilla, ja näin Pearsonin menetelmät saivat teoreettisen perustan.

[8]Eugen Slutsky (1880-1948) oli venäläinen tilastotieteilijä.

[9]Oskar Anderson (1887-1960) oli venäläis-saksalainen matemaatikko.

[10]Andrei Kolmogorov (1903-1987) oli venäläinen todennäköisyyslakentaan erikoistunut matemaatikko.

[11]Aleksandr Khinchinin (1894-1959) oli venäläinen todennäköisyysteoriaan eriskoistunut matemaatikko.

[12]Herman Wold (1908-1992) oli ruotsalainen tilastotieteilijä.

1900-luvun alussa puolestaan Udny Yule tutki auringonpilkkusarjoja ja esitteli autoregressiivisia malleja (AR), joissa satunnaishäiriöiden summien käyttäytyminen KRL:n mukaan mahdollisti harmonisten jaksojen tunnistamisen suuresta datasta huolimatta. Uudet tulokset aikasarja-analyyseissä siis rakentuivat jo kehiteltyjen asymptoottisten tuloksien pohjalta.

Myöhemmin, erityisesti 1970- ja 1980-luvuilla, asymptoottisten menetelmien merkitys kasvoi merkittävästi. Esimerkiksi Brockwellin ja Davisin klassinen teos *Time Series Theory and Methods [10]*, joka ilmestyi ensimmäisen kerran vuonna 1987, käsittelee asymptoottisia tuloksia ja keskeistä raja-arvolausetta. Myös Hamiltonin *Time Series Analysis (1994) [19]*, sekä Davidsonin ja MacKinnonin *Estimation and Inference in Econometrics (1993) [32]*, sisältävät omat kappaleensa asymptoottisista menetelmistä. Näin ollen keskeinen raja-arvolause ja asymptoottiset menetelmät aikasarja-analyyseissä nousivat tarkemman tutkimuksen kohteiksi vasta 1970- ja 1980-luvuilla, kun menetelmät olivat kehittyneet pidemmälle ja soveltamiseen oli enemmän työkaluja. Tämä kehityksen viive on luonnollinen, sillä aikasarja-analyysissä keskeinen kysymys on havaintojen välinen riippuvuus, joka on käsitteellisesti ja matemaattisesti huomattavasti monimutkaisempi ilmiö kuin riippumattomuus. Riippuvuuden muotoja on käytännössä äärettömän paljon erilaisia, minkä vuoksi KRL:ää koskien tulosten muotoileminen ja todistaminen on ollut selvästi haastavampaa kuin riippumattomille, mutta ei samoinjakautuneille satunnaismuuttujille annetut vastaavat tulokset.

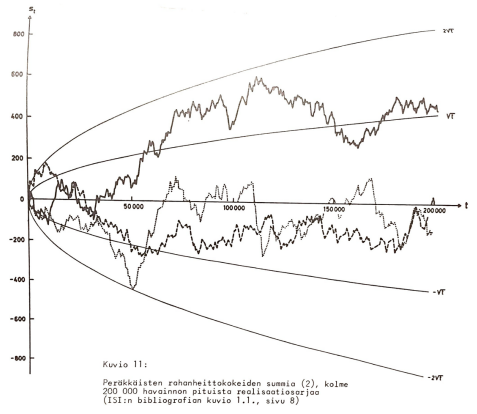
### 3.2.3 Leo Törnqvist ja Woldin esimerkin soveltaminen

Leo Törnqvist (1911-1983) oli maailmalla tunnettu aikasarjoihin erikoistunut suomalainen tilastotieteilijä. Hänet valittiin Helsingin yliopiston valtiotieteellisen tiedekunnan tilastotieteen professorin virkaan vuonna 1950. Törnqvist muun muassa kunnostautui teoreettisen tilastotieteen termien kääntämisessä englannin kielestä suomen kielelle. Näitä käännöksiä olivat muun muassa harha, ryväsotanta ja tarkentuvuus (*eng. Consistency*) ja tyhjentävä tunnusluku (*eng. Sufficient statistic*) *[5]*.

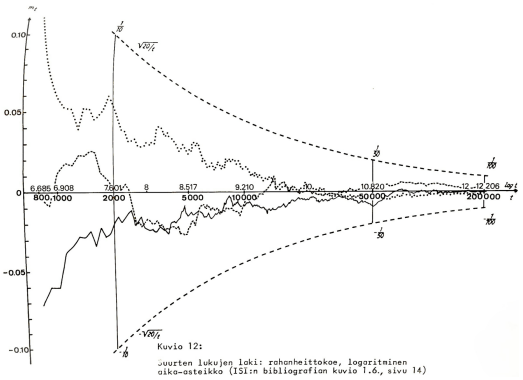
Vuonna 1974 valmistuneessa kirjassaan *Aikasarjojen analyysi ja ennustaminen* Törnqvist käsittelee muun muassa asymptoottisia tuloksia. Vertailukohdaksi hän ottaa Herman Woldin toimittamassa *Bibliography on time series and stochastic processes* teoksessa esitetyn kolikonheittokokeen, joka oli simuloitu sen aikaisilla tietokoneilla. Kokeessa oli simuloitu stokastisia prosesseja 200 000 kertaa ja tästä oli piirretty realisaatiota vastaavat aikasarjat. Koe etenee kirjassa (*[39] s. 47-52*) Törnqvistin omia sanoja mukaillen seuraavasti: Olkoon satunnaismuuttujien jono $s_1, s_2, ...,$, jonka jäsenet $s_t$ on määritelty peräkkäisinä summina $s_t = \xi_1 + \xi_2 + \cdots + \xi_t$, jossa $t = 1, 2, ...$, ja jossa summan termit $\xi_1, \xi_2, ...$, ovat satunnaismuuttujia, joilla on seuraavat ominaisuudet: Kaikki $\xi_1, \xi_2, ...$ ovat keskenään riippumattomia ja noudattavat samaa jakaumaa, jonka kertymäfunktio on $F(x)$. Jos annamme pelkän $\xi$:n edustaa jokaista $\xi_t$:tä, on siis voimassa $P(\xi \leq x) = F(x)$. Muuttujien $\xi_1, \xi_2, ...$ odotusarvo on nolla ja varianssi yksi, eli $E(\xi) = 0$ ja $\sigma^2(\xi) = 1$. Tämän jälkeen Törnqvist ohjastaa kirjassa, että peräkkäisten summien jono (1), jonka (2)-(3) ja ominaisuudet a-c määrittelevät, muodostaa stokastisen prosessin. Hän sanoo, että esimerkissä $\xi_t$ on kaksiarvoinen satunnaismuuttuja, joka voi saada arvot 1 ja -1 ja kummankin todennäköisyydellä $\frac{1}{2}$. Kyseessä on siis klassinen rahanheittokoe, jossa esimerkiksi 1 edustaa kruunaa ja -1 klaavaa.

Hän viittaa alla olevaan kuvaan 9, jossa esitetään kolme peräkkäisten summien (2)

realisaatiota. Kukin realisaatio muodostaa sarjan $Y = s_t$, jossa $t = 1, 2, ..., 200\,000$, Törnqvist tarkentaa, että nämä on laskettu tietokoneen simuloimasta 200 000 rahanheiton sarjasta. Hän jatkaa kertomalla, että kuvaan 9 piirretyt paraboliset käyrät 4a ja 4b ovat yhtälöitä $Y = \pm\sqrt{t}$ ja $Y = \pm 2\sqrt{t}$. Oletuksista a-c seuraa, että summan (2) varianssi on $t$ ja $E(s_t^2) = t$ kaikilla $t = 1, 2, ...$, ja näin ollen käyrä (4a) ilmaisee summan $s_t$ keskihajonnan kasvua $\sqrt{t}$:n mukaan.




Kuva 9: Peräkkäisten rahanheittokokeiden summia (2), kolme 200000:n havainnon pituista realisaatiosarjaa. Ks, (*[43]*, s. 49.)




Kuva 10: Suurten lukujen laki: Rahanheittokoe, logaritminen aika-asteikko. Ks, (*[43]*, s. 51.)

Keskeisen raja-arvolauseen mukaan summan $s_t$ jakauma lähestyy normaalijakaumaa, kun $t$ lähestyy kohti ääretöntä. Toisin sanoen

$$\lim_{t\to\infty} P(|s_t| < \lambda\sqrt{t}) = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\, dx. \tag{2}$$

Kuvan 9 käyrät $Y = \pm\lambda\sqrt{t}$ on piirretty $\lambda$:n arvoilla 1 ja 2. Yhtälön 2 perusteella on siis $P(|s_t| < \sqrt{t}) = 0.68$, kun $t$ on suuri. Kun $\lambda = 2$ on vastaava todennäköisyys 0.95. Merkitään $m_t$:llä satunnaismuuttujien $\xi_t$ peräkkäisiä keskiarvoja

$$m_t = \frac{s_t}{t} = \frac{\xi_1 + \cdots + \xi_t}{t}, \quad t = 1, 2, ...$$

Kuvasta 10 nähdään kuvan 9 kolmea realisaatiojonoa vastaavat keskiarvojonot $m_t$. Kuvan 10 logaritminen t-asteikko kiinnittää huomion suurten $t$-arvojen realisaatioihin ja tuo niiden rajakäyttäytymisen näkyviin selvemmin kuin lineaarinen asteikko. Kuvasta 10 nähdään, että kun $t$ on suuri, esimerkiksi 200 000 tai 50 000 keskiarvo $m_t$ on likimäärin sama kuin odotusarvo $E(\xi)$, eli $m_t \approx E(\xi) = 0$ Kun $t = 200\,000$, kaikki kolme realisaatiota ovat välissä $-1/100 < m_{200\,000} < 1/100$ ja vastaavasti, kun $t = 50000$, niin ne ovat välissä $-1/50 < m_{50\,000} < 1/50$. Kuva 10 havainnollistaa itse asiassa suurten lukujen lakia, jonka mukaan mille hyvänsä annetulle $\varepsilon < 0$ pätee

$$\lim_{t\to\infty} P(|m_t - E(\xi)| < \varepsilon) = 1.$$

Vaikka koe ei suoranaisesti liity varsinaisiin aikasarjoihin, koska tässä summan termit ovat toisistaan riippumattomia ja samoin jakautuneita, niin Törnqvistin kirjan (*[39] s. 47-52*) esittämä kohta kuitenkin osoittaa, että asymptoottisia tuloksia ja keskeiseen raja-arvolauseeseen perustuvia menetelmiä tutkittiin myös suomalaisessa aikasarja-analyysiä käsittelevässä tiedekirjallisuudessa jo 1970-luvulla. Woldin tutkielmat puolestaan olivat 1960-luvulta, jolloin kolmannen sukupolven tietokoneet saapuivat tutkijoiden käyttöön. Tietokoneiden kehittymisen myötä KRL:n menetelmiä aikasarja-analyyseissä alettiin tutkimaan yhä enemmän.

## 3.3 Jarl Waldemar Lindeberg - suomalainen KRL:n kehittäjä

Kuva 11: Jarl Waldemar Lindeberg *[33]*.

Jarl Waldemar Lindeberg (1876-1932) oli suomalainen matemaatikko, joka sai tunnustusta erityisesti keskeisen raja-arvolauseen kehittäjänä ja niin sanotusta Lindebergin ehdosta (*Eng. Lindeberg condition*), jossa perinteisestä KRL:stä poiketen ei tarvita identtisesti jakautuneita muuttujia. Vaikka tämä ehto ei tässä työssä ole relevantti, sillä työssäni tutkitaan aikasarjojen kautta sellaisia KRL:n tapauksia, joissa havainnot eivät ole riippumattomia, niin halusin kuitenkin tuoda esiin merkittävän suomalaisen KRL:än kehitykseen vaikuttaneen matemaatikon, joka on saanut laajaa kansainvälistä huomiota ja hänet mainitaan lukuisissa tieteellisissä kirjoissa sekä artikkeleissa. Lisäksi työ liittyy vahvasti keskeisen raja-arvolauseen menetelmiin, joten tämä oli hyvä tilaisuus tuoda esiin hänen merkittävää työtä aiheen parissa.

Lindeberg syntyi Helsingissä ja hänen vanhempansa olivat kanslianeuvos Karl Leonard Lindeberg ja Olga Katarina Hallonblad. Lindeberg valmistui ylioppilaaksi Helsingin ruotsalaisesta "Läroverket för gossar och flickor" -lukiosta vuonna 1893. Tämän

jälkeen hän lähti opiskelemaan Helsingin yliopistoon ja valmistui filosofian kandidaatiksi ja maisteriksi vuonna 1897, sekä filosofian lisensiaatiksi vuonna 1901. Professorin arvonimen hän sai vuonna 1919 *[42]*.

Lindeberg oli lahjakas matematiikassa jo hyvin nuorena ja hän tiedosti kykynsä myös itse. Valmistuttuaan maisteriksi hän lähti vuodeksi opiskelemaan Pariisiin, josta hän sai myös väitöskirjansa aiheen koskien osittaisdifferentiaaliyhtälöitä. Pariisin vuoden jälkeen Lindeberg vietti vuoden perheensä kanssa Kofulla, jonka jälkeen vuonna 1900 hän väitteli, ja vastaväittelijänä toimi Ernst Lindelöf.[13] 1902 hänet nimitettiin dosentiksi ja 1905 adjunktiksi.[14] Professorin arvonimi hänelle myönnettiin vuonna 1919, mutta hän piti varsinaisista opetustöistä niin paljon, ettei koskaan tavoitellut varsinaista professuuria. Hän myös opetti teknillisessä korkeakoulussa vuosina 1911-1918. Lindeberg työskenteli hyvin laajasti ja hänen tutkimuksensa suuntautuivat muun muassa osittaisdifferentiaaliyhtälöihin, variaatiolaskentaan ja funktioteoriaan. Viimeiset vuotensa hän omisti todennäköisyyslaskennalle ja tilastotieteelle, aloille, joista hän tuli tunnetuksi. Innoituksensa hän luultavasti sai toimiessaan lyhyen aikaa eräässä henkivakuutusyhtiössä, jossa hänen päätavoitteenaan oli yhtiön vakavaraisuuden selvittäminen. Myöhemmin hän myös toimi aktiivisena jäsenenä Suomen aktuaariyhdistyksessä *[20]*.

Vuonna 1922 Lindeberg julkaisi kuuluisan artikkelinsa *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung*, joka julkaistiin saksalaisessa matematiikan aikakausilehdessä *Mathematische Zeitschriftenissä [6]*, jossa sivuilla 211-225 käsitellään Lindebergin artikkelia. Tässä artikkelissa Lindeberg esittää ehdon, jonka toteutuessa keskeinen raja-arvolause pätee riippumattomille, mutta ei välttämättä samoin jakautuneille satunnaismuuttujille, joilla on äärelliset varianssit. Tämä lause katsotaan nykyään klassisen keskeisen raja-arvolauseen lopulliseksi muodoksi *[33]*. Huomionarvoista on myös, että vuonna 1934 kuuluisa englantilainen matemaatikko Alan Turing kehitti saman ehdon, mutta kuuli sitten, että suomalainen matemaatikko oli jo kehittänyt sen 12 vuotta aikaisemmin *[9]*. Tämä kaiketi johtui siitä, että tuohon aikaan tiedonkulku oli huomattavasti hitaampaa kuin tänä päivänä.

Vielä tietämättään Lyapunovin aikaisemmista töistä Lindeberg oli vuonna 1920 todistanut KRL:n normitetuille summille seuraavasti

$$\sum_{k=1}^{n} \frac{X_k}{r_n},$$

jossa $X_k$ ovat toisistaan riippumattomia satunnaismuuttujia, joilla kullakin on jakauma $U_k$, jossa $\mu = 0$ ja varianssi $\sigma_k^2$, sekä äärellinen kolmannen kertaluvun absoluuttinen momentti, olettaen että

$$\frac{1}{r_n^3} \sum_{k=1}^{n} \int_{-\infty}^{\infty} |x|^3 dU_k(x) \to 0 \quad (n \to \infty), \quad r_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2}.$$

*[14]*. Lindeberg muokkasi tätä Lyapunovin aikaisemmin kehittelemää lausetta edelleen ja lievensi sen vaatimia ehtoja, kun taas Lyapunovin versio keskittyi tarkastelemaan

[13] Helsingin yliopiston matematiikan professori vuosina 1903-1938.

[14] Vanha nimitys lehtorista/apulaisprofessorista.

suurten poikkeamien vaikutusta vahvalla ehdolla, joka rajoitti yksittäisten satunnaismuuttujien suurten poikkeamien vaikutusta summan käyttäytymiseen. Lyapunovin ehto oli oikea, mutta Lindebergin ehto on välttämätön ja riittävä. Hän teki siitä muutamia erilaisia versioita, mutta kenties kuuluisin on seuraava alla esitetty "Theorem III" (*[14] s. 233*), (*[6] s. 219-220*), jonka hän julkaisi vuonna 1922.

Olkoon $U_1, U_2, ..., U_n$ jakaumafunktiot $n$:lle toisistaan riippumattomalle satunnaismuuttujalle $u_1, u_2, ..., u_n$, joilla jokaisella on odotusarvo 0 ja varianssi $\sigma_k^2$, jossa $\sum_{k=1}^{n} \sigma_k^2 = 1$. Olkoon

$$U(x) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} U_n(x - t_1 - t_2 - \cdots - t_{n-1}) dU_{n-1}(t_{n-1}) \cdots dU_1(t_1).$$

Tällöin $U$ on kaikkien satunnaismuuttujien summan jakauma. Olkoon

$$s(x) := \begin{cases} |x|^3, & \text{jos} \quad |x| < 1, \\ x^2, & \text{muutoin.} \end{cases}$$

Tällöin, vaikka positiivinen luku $\varepsilon$ valitaan mielivaltaisen pieneksi, voidaan aina valita positiivinen kokonaisluku $N$ siten, että kaikilla $n \geq N$ pätee

$$\left| U(x) - \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \right| < \varepsilon,$$

jos

$$\sum_{k=1}^{n} \int_{-\infty}^{\infty} s(x) d, \quad U_k(x) < n.$$

(*[14] s. 233*), (*[6] s. 219-220*).
Tämä voitaisiin lyhyesti esittää KRL:n tunnetussa nykyisessä standardoidussa muodossa

$$\frac{S_n}{\sqrt{Var(S_n)}} \to \mathcal{N}(0, 1),$$

jossa

$$S_n = \sum_{k=1}^{n} u_k, \; E(u_k) = 0 \text{ ja } Var(u_k) = \sigma_k^2, \text{ jossa } \sum_{k=1}^{n} \sigma_k^2 = 1.$$

Koska varianssien summa on 1, niin $\sqrt{Var(S_n)} = 1$, ja yhtälö yksinkertaistuu muotoon

$$S_n \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{kun } n \to \infty.$$

Mikäli kyseessä olisi samoin jakautuneet ja toisistaan riippumattomat satunnaismuuttujat $X_n$, jossa $X_k$:t olisivat samoin jakautuneita ja toisistaan riippumattomia, niin varianssien summa kasvaisi lineaarisesti $\sum_{k=1}^{n} \sigma_k^2 = n$, jolloin keskiarvoa $\frac{1}{n} \sum_{k=1}^{n} X_k$ pitäisi skaalata tekijällä $\sqrt{n}$, jotta kokonaisvarianssi asettuisi arvoon 1. Tällöin esitys olisi

$$\sqrt{n} \left( \frac{1}{n} \sum_{k=1}^{n} X_k - E(X_k) \right) / \sigma \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{jossa } n \to \infty.$$

Tästä lisää seuraavassa luvussa.

Lindebergin lause ja hänen työnsä tilastotieteen, asymptoottisten menetelmien ja todennäköisyyslaskennan parissa olivat uraauurtavia. Hän jakoi Lyapunovin työtä ja uusilla havainnoilla saattoi loppuun keskeisen raja-arvolauseen menetelmien kehittämisen, kuten ne tänä päivänä tunnemme. Kuten kirjassa *A History of the Central Limit Theorem [14]* sanotaan

*"The complete mathematical work of Lindeberg contains only one truly outstanding, virtually epochal performance: the proof of the CLT under a very weak condition, which under certain "natural" assumptions even proved to be necessary. Lindeberg's arguments were based on an entirely new analytic method, which would later be applied to far more general problems."*

Lindeberg teki myös muita kirjallisia julkaisuja ja eräs hänen kuuluisa teoksensa on vuonna 1927 ilmestynyt *Todennäköisyyslasku ja sen käytäntö tilastotieteessä. Alkeellinen esitys*, joka oli ensimmäinen tilastotiedettä käsittelevä oppikirja Suomessa. Kirja sai alkunsa siitä, kun Suomen Tilastoseuran tilastotiedettä käsittelevälle kurssille ilmoittautui vain 8 henkilöä, jota ei pidetty riittävänä. Näin Suomen Tilastoseura ja muiden tieteellisten seurojen valiokunta kysyi Lindebergiä tilastotiedettä käsittelevän oppikirjan tekijäksi, johon Lindeberg suostui *[38]*.

Valitettavasti tämän loistavan matemaatikon maallinen taivallus loppui liian lyhyeen, sillä Jarl Waldemar Lindeberg menehtyi vuonna 1932, vain 56-vuotiaana. Hänet tunnettiin ahkerana ja inspiroivana työkaverina. Erittäin tunnettu hän oli myös hyvästä huumoristaan. Kuten kirja *Statisticians of the Centuries [20]* sanoo

*"He spoke with much modesty and joking irony of his own relationship to science. You see: he once said, owning a country place with forest and farming land, in Helsinki I can defend my laziness by saying that I am really a farmer, and in the country by claiming really to be a scientist"*

# 4 Keskeisen raja-arvolauseen perusoletukset

Seuraavassa pääosin lähteeseen *[25]* nojaten esitetään lyhyesti KRL:n perusoletukset, sekä standardoitu ja skaalattu muoto.

## 4.1 Klassinen keskeinen raja-arvolause (KRL)

KRL sanoo, että kun otoskoko $n$ kasvaa, niin tiettyjen oletusten vallitessa otoskeskiarvon jakauma lähestyy normaalijakaumaa riippumatta havaintojen alkuperäisestä jakaumasta. Olkoon $X_i$, jossa $i = 1, 2, ..., n$, riippumattomia ja samoin jakautuneita satunnaismuuttujia, joilla $E(X_i) = \mu$ ja $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Tällöin keskeinen raja-arvolause voidaan esittää kahdessa eri muodossa sen mukaan, standardoidaanko otoskeskiarvo vai ei. Standardoidussa muodossa saadaan

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{kun } n \to \infty, \tag{3}$$

kun taas standardoimatonta muotoa käyttäen tulos voidaan kirjoittaa muodossa

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \text{kun } n \to \infty. \tag{4}$$

Kaava (3) standardoi summan yksiköihin $\sigma$, jolloin jakauma lähestyy normaalijakaumaa, kun $n \to \infty$. Kaava (4) puolestaan säilyttää alkuperäisen mittakaavan, jolloin normaalijakauman varianssi on $\sigma^2$. Kummassakin tapauksessa otoskeskiarvo

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

on otettu riippumattomista ja samoinjakautuneista satunnaismuuttujista $X_i$.

### 4.1.1 Standardoitu muoto

Keskeinen raja-arvolause sanoo, että otoskeskiarvojen jakauma lähestyy normaalijakaumaa, kun otoskoko kasvaa. Yksi tavallisimmista tavoista ilmaista tämä on standardoida otoskeskiarvon poikkeama todellisesta odotusarvosta. Tämä johtaa tapaukseen (3), jossa otoskeskiarvo on standardoitu, eli siitä on vähennetty todellinen odotusarvo $\mu$ ja jaettu keskihajonnalla $\sigma/\sqrt{n}$, jolloin saadaan muuttujajono, joka konvergoi standardinormaalijakaumaan $\mathcal{N}(0, 1)$.

Tätä standardoitua muotoa käytetään laajasti tilastollisessa päättelyssä, erityisesti hypoteesien testaamisessa ja luottamusväleissä. Hypoteesien testaamisessa se toimii testisuureena, jonka avulla verrataan havaittua keskiarvoa nollahypoteesin mukaiseen arvoon. Esimerkiksi testattaessa $H_0 : \mu = \mu_0$, lasketaan testisuure

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma},$$

ja tätä verrataan standardinormaalijakauman kriittisiin arvoihin.

Likimääräisen luottamusvälin tapauksessa kaava $(\bar{X} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n})$ antaa approksimatiivisen $100(1-\alpha)\%$:n luottamusvälin odotusarvolle $\mu$, kun keskihajonta $\sigma$ tunnetaan. Jos $X_i \sim \mathcal{N}(\mu, \sigma^2)$, niin standardoidun otoskeskiarvon $\sqrt{n}(\bar{X} - \mu)/\sigma$ jakauma on täsmällisesti $\mathcal{N}(0, 1)$. Tämä tarkoittaa, että tulos liittyy muotoon, jossa otoskeskiarvo on standardoitu. Tämä on erityisen hyödyllinen esimerkiksi luottamusväleissä, koska se mahdollistaa vertailun suoraan standardinormaalijakaumaan.

### 4.1.2 Skaalattu muoto

Tapauksessa (4) varianssi skaalautuu otoskoon mukaan $Var(\bar{X}) = \frac{\sigma^2}{n}$. Kun otoskeskiarvosta vähennetään odotusarvo $\mu$ ja kerrotaan tekijällä $\sqrt{n}$, saadaan muuttuja

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \text{kun } n \to \infty.$$

Standardoimalla vielä jakamalla $\sigma$:lla saadaan

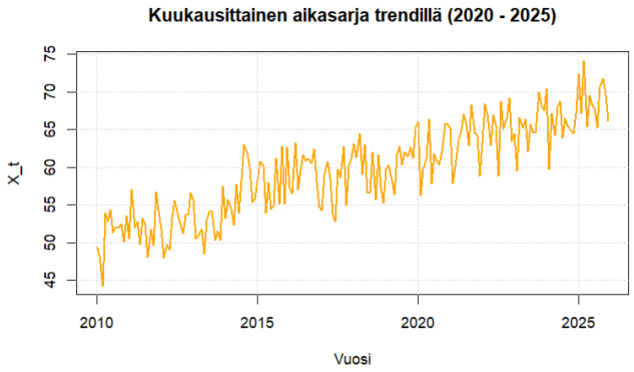
$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1), \text{ kun } n \to \infty.$$

Tämä tarkoittaa, että otoskeskiarvo $\bar{X}$ säilyttää odotusarvon $\mu$ ja sen varianssi pienenee $\sigma^2/n$ otoskoon kasvaessa. Erityisesti, jos $X_i \sim \mathcal{N}(\mu, \sigma^2)$, niin otoskeskiarvo $\bar{X}$ on täsmällisesti normaalijakautunut

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

# 5 Aikasarjoista ja niiden stationaarisuudesta

Tässä osiossa havainnollistetaan yksinkertaisessa muodossa aikasarjojen perusolemus ja stationaarisuus niissä. Aikasarjoilla tarkoitetaan joukkoa peräkkäisiä havaintoja $X_1, X_2, ..., X_t$, jotka kaikki ovat kirjattu tietyllä ajanhetkellä $t$.




Kuva 12: Kuukausittainen aikasarja trendillä

Aikasarja etenee ajan suhteen pisteestä toiseen, ja sitä voidaan havainnollistaa käyränä, jossa kukin piste kuvaa havaintoa tietyllä aikavälillä. Kyseessä on siis muuttuja, jota seurataan ja tallennetaan säännöllisin aikavälein, kuten esimerkiksi osakekurssit, säätiedot, työttömyysaste tai sähkönkulutus. Kuvassa 12 on esitetty kuvitteellinen kuukausittainen aikasarja, joka kuvaa arvojen kehitystä ajanjaksolla 2010-2025. Sarjassa on yhteensä 192 havaintoa (16 vuotta $\times$ 12 havaintoa), ja jokainen havainto $X_t$ edustaa muuttujan arvoa kyseisen kuukauden kohdalla. Tämä aikasarja on simuloitu niin, että siinä yhdistyy nouseva lineaarinen trendi ja satunnaista vaihtelua. Trendin ansiosta havaintojen taso kasvaa ajan myötä, mutta kohina tekee aikasarjasta epäsäännöllisen ja realistisen. Tällaisia sarjoja voi syntyä esimerkiksi, kun seurataan energian kulutusta, joka kasvaa ajan myötä, mutta vaihtelee kuukausittain sään tai kulutustottumusten mukaan.

Aikasarjojen analysointiin vaikuttavat muun muassa trendi, eli pitkän aikavälin kehitys, kausivaihtelut, eli säännönmukaisesti toistuvat jaksot kuten vuodenajat, sekä satunnaiskomponentit, jotka kuvaavat ennustamatonta vaihtelua. Kuten kaikessa tilastollisessa mallintamisessa, myös aikasarja-analyysi aloitetaan tutustumalla käytettävään dataan. Aluksi tarkastellaan muun muassa havaintojen määrää, aikaväliä, yksiköitä ja keskeisiä tunnuslukuja, kuten keskiarvoa, mediaania, minimiä, maksimiarvoa ja hajontaa. Lisäksi selvitetään mahdolliset puuttuvat arvot sekä datan laatu. Aikasarjoissa on myös tärkeää tarkistaa, esiintyykö sarjassa trendiä tai kausivaihtelua, pysyykö varianssi vakiona (stationaarisuus) ja onko havaittavissa poikkeamia.

Seuraavaksi esitetään yleisimmät aikasarjamallit. Nämä mallit esitellään lähes poikkeuksetta alan kirjallisuudessa ennen laajempia yleistyksiä, ja niitä on käytetty myös tässä työssä analyysin lähtökohtana.

## 5.1 AR($p$)-prosessi

Tässä AR($p$)-prosessia käsittelevässä osiossa päälähteenä on käytetty lähteitä *[34]* ja *[16]*, sekä todistuksessa lähdettä *[2]*.

### 5.1.1 AR($p$)-prosessin määritelmä

AR($p$) on aikasarjamalli, jossa nykyinen havainto riippuu $p$ aikaisemmasta havainnosta. Malli voidaan esittää muodossa

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \tag{5}$$

jossa $X_t$ on aikasarjan arvo hetkellä $t$. Mallin autoregressiiviset kertoimet ovat $\phi_1, \phi_2, ..., \phi_p$, jossa $\phi_p \neq 0$, ja jotka kerrotaan viivästetyillä havainnoilla $X_{t-1}, X_{t-2}, ..., X_{t-p}$ ja $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ on normaalisti jakautunut virhetermi, joka muodostaa valkoista kohinaa. Mallissa (5) satunnaismuuttujan $X_t$ odotusarvo on nolla. Mikäli odotusarvo $\mu$ on jokin muu kuin nolla ($\mu \neq 0$), voidaan malli esittää seuraavassa muodossa

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t. \tag{6}$$

*[34].* AR(1)-prosessissa varianssi voidaan ilmaista suljetussa muodossa $Var(X_t) = \frac{\sigma^2}{1-\phi_1^2}$. AR(2)-prosessin varianssin (21) johtaminen edellyttää tarkempia määritelmiä, sekä niin kutsuttua Yule-Walkerin menetelmää ja siihen perustuvia yhtälöitä. Nämä yhtälöt perustuvat autokovarianssien ja autokorrelaatioiden viiveoperaattoreihin. Stationaarisessa AR(p)-prosessissa autokovarianssi ja autokorrelaatio viiveellä $k > 0$ voidaan laskea rekursiivisesti seuraavasti

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p} \tag{7}$$
$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p} \tag{8}$$

*[16].* Seuraavaksi Yule-Walker-yhtälöiden todistus. Oletetaan prosessin keskiarvo nollaksi. Tällöin prosessi voidaan määritellä mallin (5) mukaan, jolloin yhtälöt todistetaan seuraavasti

*Todistus.*

$$\begin{aligned}\gamma_k = Cov(X_t, X_{t-k}) &= E[X_t X_{t-k}] = E[(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t)X_{t-k}] \\ &= \phi_1 E[X_{t-1}X_{t-k}] + \phi_2 E[X_{t-2}X_{t-k}] + \cdots + \phi_p E[X_{t-p}X_{t-k}] + E[\varepsilon_t X_{t-k}] \\ &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p}\end{aligned}$$

□

*[2].* Toinen muoto (8) saadaan jakamalla yhtälön molemmat puolet $\gamma_0$:lla ja huomaamalla, että $\rho_k = \gamma_k/\gamma_0$.

### 5.1.2 AR($p$)-prosessin stationaarisuus

AR($p$)-prosessin stationaarisuutta voidaan tarkastella niin kutsutun viiveoperaattorin *(eng. Backshift operator)* ja siitä johdetun karakteristisen polynomin juurien avulla. Viiveoperaattoria käyttäen AR($p$)-prosessi (6) voidaan kirjoittaa muodossa

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(X_t - \mu) = \varepsilon_t.$$

Autoregressiivinen operaattori $B$ määritellään muodossa

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$

jolloin malli voidaan tiivistää muotoon $\phi(B)(X_t - \mu) = \varepsilon_t$. Stationaarisuusehtoa varten tarkastellaan karakteristista polynomia

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \ \phi_p \neq 0, \tag{9}$$

jossa viiveoperaattori $B$ on korvattu kompleksimuuttujalla $z$. AR($p$)-prosessi on stationaarinen vain jos kaikki polynomin (9) juuret sijaitsevat yksikköympyrän ulkopuolella, eli

$$\phi(z) = 0 \implies |z| > 1$$

*[34]*.

Vaikka myöhemmin työssä esitettävät AR(1)- ja AR(2)-mallit ovat erikoistapauksia yleisestä AR($p$)-mallista, niin sen laajempaan käyttöön ei tässä työssä enää palata. Tämä johtuu siitä, että AR($p$)-mallin laajempi käyttö vaatisi huomattavasti monimutkaisempien menetelmien käyttöä, jotka vaikeuttaisivat KRL:n havainnollistamista AR($p$)-prosessissa ja tekisivät esityksestä matemaattisesti sekavan. Tämän vuoksi AR(1)- ja AR(2)-prosessit ovat riittäviä tuomaan esiin AR($p$)-KRL:n käyttäytymisen. Osiossa 7 AR(1)-prosessi käsitellään omana tapauksenaan, vaikka AR(1)-malli on AR(2)-prosessin erikoistapaus, jossa $\phi_2 = 0$, niin sen yksinkertainen rakenne tekee siitä erinomaisen lähtökohdan keskeisen raja-arvolauseen soveltamisen ja aikasarjamallien ominaisuuksien havainnollistamiseen. Monet AR(2)-mallin piirteet ja ilmiöt voidaan ensin ymmärtää intuitiivisesti AR(1)-mallin kautta, mikä helpottaa myöhempää yleistystä monimutkaisemmiksi malleiksi. Lisäksi AR(1)-malli on käytännön sovelluksissa hyvin yleinen ja tarjoaa usein riittävän tarkan kuvauksen ajallisesta riippuvuudesta yksinkertaisella parametrisaatiolla.

## 5.2 MA($q$)-prosessi

Tässä MA($q$)-prosessia käsittelevässä osiossa päälähteinä on käytetty lähteitä *[28]* ja *[41]*.

### 5.2.1 MA($q$)-prosessin määritelmä

MA($q$)-prosessi määritellään seuraavasti

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{10}$$

jossa odotusarvo $E(X_t) = \mu$. Liukuvan keskiarvon kertoimet merkitään $\theta_1, \ldots, \theta_q$, ja $\varepsilon_t$ on normaalijakautunut virhetermi, joka muodostaa valkoista kohinaa. Toisin kuin AR($p$)-prosessissa, MA($q$)-prosessin tapauksessa stationaarisuus on automaattinen, koska prosessi on määritelty vain äärellisen määrän kohinatermejä sisältävänä lineaarisena yhdistelmänä. Tässä mallissa nykyinen havainto $X_t$ riippuu suoraan korkeintaan $q$ aikaisemmasta virhetermistä $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$. Tämä tarkoittaa, että tavallisen KRL:n riippumattomuusoletus ei siten päde. Prosessin varianssi puolestaan saadaan seuraavasti $Var(X_t) = \gamma_0 = \sigma^2(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)$.

MA($q$)-prosessien tärkeä ominaisuus on niiden rajoittunut riippuvuusrakenne. Vaikka $\varepsilon_t$:t ovat riippumattomia, prosessin $X_t$ havainnot voivat olla korreloituneita, mutta

ainoastaan korkeintaan $q$ viiveen verran. Tämä ilmenee prosessin autokovarianssifunktioissa

$$\begin{aligned}\gamma_0 &= \sigma^2(\theta_0^2 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\\ \gamma_1 &= \sigma^2(\theta_1\theta_0 + \theta_2\theta_1 + \cdots + \theta_q\theta_{q-1})\\ \gamma_k &= 0,\ |k| > q.\end{aligned} \tag{11}$$

Vastaavasti autokorrelaatiofunktio on muotoa

$$\rho_k = \frac{\gamma_k}{\gamma_0}, \tag{12}$$

ja se katkeaa tarkalleen viiveen $q$ kohdalla. Tämä tarkoittaa, että MA($q$)-prosessin autokorrelaatiokaavio (ACF) sisältää ei-nolla-arvoja vain viiveillä $h = 0, 1, \ldots, q$, ja on nolla kaikilla tätä suuremmilla viiveillä *[28]*.

### 5.2.2 Kääntyvyysehto MA($q$)-prosesseissa

Vaikka MA($q$)-prosessi on aina stationaarinen, sille asetetaan käytännössä usein lisäehto, jota kutsutaan kääntyvyysehdoksi *(eng. Invertibility)*. Kääntyvyysehdolla tarkoitetaan sitä, että MA($q$)-prosessi voidaan esittää yksikäsitteisesti äärettömänä autoregressiivisena AR($\infty$)-mallina. Ilman kääntyvyyttä prosessi on moniselitteinen, ja parametreja ei voida identifioida yksikäsitteisesti.

Kääntyvyyttä voidaan tarkastella karakteristisen juurten polynomien avulla

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q, \tag{13}$$

jossa kaikkien juurten tulee sijaita yksikköympyrän ulkopuolella

$$\theta(z) = 0 \implies |z| > 1.$$

Kääntyvyysehdot ovat tärkeitä MA($q$)-prosessien yksikäsitteisen esityksen kannalta, mutta ne eivät ole MA($q$)-prosessin edellytyksiä. MA($q$)-prosessi säilyy stationaarisena ja $q$-riippuvaisena, vaikka kääntyvyysehdot eivät täyttyisi, kunhan kohinatermit $\varepsilon_t$ ovat äärellisen varianssin omaavia. Tämän seurauksena KRL pätee edelleen ja otoskeskiarvon jakauma lähestyy normaalijakaumaa, vaikka kääntyvyysehtoja rikottaisiin. Esimerkiksi MA(1)-prosessin polynomi on muotoa $\theta(z) = 1 + \theta_1 z$. Kääntyvyysehdoksi saadaan tällöin

$$1 + \theta_1 z = 0 \implies \left|-\frac{1}{\theta_1}\right| > 1 \implies |\theta_1| < 1 \implies -1 < \theta_1 < 1. \tag{14}$$

*[41]*. Jos ehto ei täyty, prosessia voidaan edelleen käyttää, mutta sen AR($\infty$)-esitystä ei enää ole olemassa, eivätkä parametrit ole yksiselitteisesti identifioituvia.

## 5.3 ARMA($p, q$)-prosessi

Tässä osiossa käsitellään ARMA($p, q$)-mallia ja päälähteenä on käytetty Robert H. Shumwayn ja David S. Stofferin teosta *Time Series Analysis and Its Applications with R Examples [34]*.

### 5.3.1 ARMA$(p, q)$-prosessin määritelmä

ARMA$(p, q)$-malli, eli yhdistetty autoregressiivinen liukuvan keskiarvon malli (eng. Autoregressive moving average model) yhdistää AR$(p)$- ja MA$(q)$-mallit. Tämä tuo malliin joustavuutta, varsinkin tilanteissa, joissa pyritään mallintamaan monimutkaisempia aikasarjatilanteita.

ARMA$(p, q)$-prosessi määritellään seuraavasti

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

jossa $X_t$ on arvo hetkellä $t$. AR$(p)$-prosessin autoregressiiviset kertoimet ovat $\phi_1, ..., \phi_p$, jossa $\phi_p \neq 0$. MA$(q)$-prosessin liukuvan keskiarvon kertoimet ovat $\theta_1, ..., \theta_q$, jossa $\theta_q \neq 0$. Mikäli prosessin odotusarvo on nolla, niin se voidaan tällöin esittää myös MA$(\infty)$-prosessina seuraavasti

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}.$$

Mikäli $X_t$:lla on nollasta poikkeava keskiarvo $\mu$, malli voidaan esittää kahdella eri tavalla riippuen siitä, keskitetäänkö se odotusarvon $\mu$ ympärille, vai lisätäänkö $\mu$ erikseen yhtälöön, jolloin ne voidaan kirjoittaa seuraavasti

$$X_t = \mu + \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

joka on yhtäpitävä myös seuraavan muodon kanssa

$$X_t = \mu(1 - \phi_1 - \cdots - \phi_p) + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Kun $q = 0$ mallia kutsutaan asteen $p$ autoregressiiviseksi malliksi, AR$(p)$. Kun $p = 0$ mallia kutsutaan asteen $q$ liukuvan keskiarvon malliksi MA$(q)$. Mikäli $p \neq 0$ ja $q \neq 0$, kyseessä on yhdistetty ARMA-malli *[34]*.

### 5.3.2 ARMA$(p, q)$-mallin karakteristiset polynomit

Tässä tiivistetään aikaisemmissa osioissa käsitellyt AR$(p)$- ja MA$(q)$-prosessien karakteristiset polynomit (9) ja (13), jotka osaltaan vaikuttavat ARMA$(p, q)$-mallin toimivuuteen. Kuten aikaisemmin todettiin karakteristiset polynomit liittyvät stationaarisuuteen ja kääntyvyyteen. Tässä stationaarisuus 5.1.2 riippuu AR$(p)$-polynomin $\phi(z)$ juurista, kun taas kääntyvyysehto 5.2.2 määräytyy MA$(q)$-polynomin juurista. Stationaarisuuden ehdoksi saadaan, että AR$(p)$-polynomin $\phi(z)$ kaikki juuret sijaitsevat yksikköympyrän ulkopuolella, eli $|z| > 1$. Kääntyvyysehto on, että MA$(q)$-polynomin $\theta(z)$ kaikki juuret sijaitsevat yksikköympyrän ulkopuolella. Nämä ehdot varmistavat, että prosessille on olemassa vakaa esitys, joko AR$(p)$- tai MA$(q)$-muodossa.

## 5.4 Stationaarisuus

Tässä osiossa käydään läpi, mitä tarkoittaa stationaarisuus aikasarja-analyyseissä. Osio perustuu pääosin lähteisiin (*[13] s.232-234*) ja *[29]* ja siinä pyritään antamaan peruskäsitys stationaarisuudesta ennen kuin siirrytään varsinaiseen aiheeseen teoriaosuudessa.

Stationaarisuus tarkoittaa, että aikasarjan tilastolliset ominaisuudet eivät muutu ajan mukana. Käytännössä tämä tarkoittaa, että aikasarja käyttäytyy samalla tavalla riippumatta siitä, millä ajanhetkellä sitä tarkastellaan. Stationaarisuudelle voi määritellä kaksi tasoa seuraavasti

### 5.4.1 Vahva stationaarisuus

Kaikkien satunnaismuuttujien yhteisjakauma pysyy täysin samana, kun prosessia siirretään ajassa eteenpäin

$$P(y_{t_1} \in A_1, \ldots, y_{t_m} \in A_m) = P(y_{t_1+k} \in A_1, \ldots, y_{t_m+k} \in A_m) \; \forall k \in Z$$

Tämä tarkoittaa, että vahva stationaarisuus edellyttää, että yhteisjakauma pysyy samana riippumatta siitä, mitkä ajankohdat $t_1, ..., t_m$ tai joukot $A_1, ..., A_m$ valitaan *[29]*, eli koko aikasarjan todennäköisyysjakauma on ajan suhteen vakio.

### 5.4.2 Heikko stationaarisuus

Tässä aikasarjan odotusarvo $E(Y_t)$ ei riipu ajasta ja autokovarianssi $Cov(Y_t, Y_{t+k})$ riippuu vain viiveestä $k$, mutta ei itse ajanhetkestä $t$.

Stationaarisesta aikasarjasta saatu lineaarisesti suodatettu sarja on myös stationaarinen. Esimerkiksi valkoisesta kohinasta $\varepsilon_t$ saadaan stationaarinen sarja, mikäli alla oleva ehto täyttyy

$$Y_t = \mu + \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}.$$

Jotta tämä sarja olisi stationaarinen, niin kertoimien $\theta_i$ tulee täyttää seuraava ehto

$$Var(Y_t) = Var\left(\sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}\right) < \infty.$$

Tämä ehto takaa, että sarjan varianssi pysyy äärellisenä ja sarja on täten heikosti stationaarinen.

Käytännössä stationaarisuutta voidaan arvioida esimerkiksi vertaamalla prosessin käyttäytymistä eri ajanhetkillä ja tarkastelemalla sen yleistä tilastollista luonnetta. Esimerkiksi autokorrelaatiofunktiota ACF (eng. Autocorrelation function), jossa voimakas ja hitaasti hiipuva ACF saattaa viitata ei-stationaarisuuteen. Lisäksi tarkemmilla menetelmällisillä testeillä, kuten ADF-testeillä (eng. Augmented Dickey-Fuller test), jotka ovat niin sanottuja yksikköjuuritestejä *[13]*.

Tässä työssä käytettyjen aikasarjamallien AR($p$), MA($q$) ja ARMA($p, q$) stationaarisuus määräytyy osioissa 5.1–5.3 esitettyjen ehtojen mukaisesti.

# 6 $m$-riippuvaiset prosessit

KRL:n perusoletuksien mukaan jakaumat ovat samoinjakautuneita ja toisistaan riippumattomia. Aikasarjoissa riippuvuus on kuitenkin aina läsnä, jolloin klassisen KRL:n riippumattomuusoletusta rikotaan. $m$-riippuvaiset prosessit muodostavat tässä tärkeän väliluokan. Niissä riippuvuus on edelleen läsnä, mutta rajoittuu vain tiettyyn aikaväliin. Tällainen rajoitettu riippuvuus mahdollistaa KRL:n soveltamisen suhteellisen suoraviivaisesti aikasarjamalleille, joissa riippuvuus voi olla pidemmälle ulottuvaa ja vaatia vahvempia ehtoja. Luvussa nojataan erityisesti lähteeseen *[11]*, jossa käsitellään $m$-riippuvaisia prosesseja ja niiden asymptoottista käyttäytymistä.

Tässä työssä $m$-riippuvaisia prosesseja käsitellään ensisijaisesti havainnollistavana taustana, joka motivoi KRL:n soveltamista toisistaan riippuviin aikasarjoihin. Työssä ei käsitellä m-riippuvaisten prosessien KRL:n todistusta yksityiskohtaisesti, vaan varsinainen painopiste on myöhemmin tarkasteltavissa AR-, MA- ja ARMA-malleissa. Yksityiskohtaisempi käsittely ja todistukset löytyvät lähteestä (*[11] s.213*).

Lähteen (*[11] s. 212)* mukaan stokastinen prosessi $(X_t)$ on $m$-riippuvainen, jos kaikilla ajanhetkillä $t$ satunnaismuuttujajoukot $(X_j : j \leq t)$ ja $(X_j : j \geq t + m + 1)$ ovat riippumattomia. Toisin sanoen, mitkään prosessin osat, jotka ovat toisistaan enemmän kuin $m$ ajanhetkeä erillään, eivät sisällä mitään informaatiota toisistaan. Erityistapauksessa, kun $m = 0$, prosessin kaikki ajanhetket ovat toisistaan riippumattomia. Esimerkiksi aikaisemmin luvussa 5.2 esitellyt MA($q$)-prosessit ovat $m$-riippuvaisia, jossa $m = q$.

Kaikki $m$-riippuvaiset prosessit eivät kuitenkaan ole liukuvan keskiarvon muotoa (MA($q$)). Esimerkiksi Billingsley (*[8] s.364*) esittää seuraavan rakenteen. Jos $(Y_t)$ on iid-prosessi ja $f$ on reaalinen funktio, voidaan määritellä uusi prosessi $(X_t)$ asettamalla

$$X_t = f(Y_t, Y_{t+1}, \ldots, Y_{t+m}).$$

Tällöin prosessi $(X_t)$ on $m$-riippuvainen. Riippuvuus johtuu siitä, että $X_t$ ja $X_{t'}$ jakavat satunnaismuuttujia $Y_s$ vain silloin, kun $|t - t'| \leq m$. Tämä prosessi ei kuitenkaan ole MA($q$)-prosessi, koska MA-prosesseissa $X_t$ esitetään menneisyyden virhetermien funktiona $(\varepsilon_{t-j})$, kun taas tässä $X_t$ riippuu tulevaisuudesta $(Y_{t+1})$ ja voi luoda epälineaarisia riippuvuuksia havaintojen välille. MA-prosessissa riippuvuudet ovat aina lineaarisia. Mikäli funktio $f$ olisi kuitenkin lineaarinen ja

$$f(Y_n, ..., Y_{n+q}) = \sum_{i=0}^{q} \theta_i Y_{n+i}$$

ja

$$Y_n = \varepsilon_n \sim \mathcal{N}(0, \sigma^2),$$

jossa $\varepsilon_n$ on valkoista kohinaa, mikä koostuu itsenäisistä ja samoinjakautuneista satunnaismuuttujista, joiden odotusarvo $E(\varepsilon_n) = 0$ ja varianssi $Var(\varepsilon_n) = \sigma^2$, niin tällöin $X_n$ olisi $q$-riippuvainen, jossa $X_t$ riippuisi tulevaisuuden satunnaismuuttujista $Y_{t+1}, Y_{t+2}, ..., Y_{t+q}$. Tämä ei silti myöskään olisi MA($q$)-prosessi, koska se ei olisi kausaalinen ja valkoisen kohinan termit keskittyisivät tulevan ennustamiseen. Perinteinen MA($q$)-prosessi (10) puolestaan riippuu nykyhetken ja menneisyyden virhetermeistä, eli se on kausaalinen ja siinä summataan nykyhetken ja menneiden aikapisteiden termejä. Tämä on tärkeää, jos halutaan ennustaa tulevaisuutta nykyisyyden ja

menneisyyden avulla. Tätä käsiteltiin aikaisemmin luvussa 5.2 ja siihen palataan vielä myöhemmin luvussa 9.

Yksinkertainen esimerkki $m$-riippuvaisuudesta, joka ei ole MA($q$)-prosessi, voisi olla seuraava. Olkoon $X_t = Y_t Y_{t+1},$ jossa $Y_t \sim$ i.i.d. $\mathcal{N}(0,1)$. Tällöin $X_t$ ja $X_{t+2}$ eivät jaa yhteisiä satunnaismuuttujia ja ovat siten riippumattomia, mutta $X_t$ ja $X_{t+1}$ riippuvat molemmat muuttujasta $Y_{t+1}$. Tällöin prosessi on 1-riippuvainen, ja siitä otetusta itseisarvon logaritmista saadaan

$$L_t = \log|Y_t Y_{t+1}| = \log|Y_t| + \log|Y_{t+1}|,$$

jossa $L_t = \varepsilon_t + \varepsilon_{t+1}$, missä $\log(Y_t) = \varepsilon_t$. Tämä muistuttaa MA(1)-prosessia, mutta ei kuitenkaan ole perusoletuksiltaan MA(1)-prosessi, koska tässä valkoisen kohinan termit riippuisivat tulevaisuuden satunnaismuuttujasta $\varepsilon_{t+1}$, ja niiden odotusarvo ei välttämättä ole nolla ja ne eivät välttämättä jakaudu normaalisti, vaikka olisivatkin i.i.d. -muuttujia. Sen sijaan MA(1)-prosessi riippuisi menneisyyden satunnaismuuttujista $\varepsilon_t + \theta\varepsilon_{t-1}$, jossa valkoisen kohinan termit jakautuisivat normaalisti ja niiden odotusarvo olisi nolla.

Nämä esimerkit osoittavat, että $m$-riippuvuus ei rajoitu pelkästään MA($q$)-prosessien rakenteeseen, vaan sitä voi ilmetä myös muissa, ei-lineaarisissa tai tulevaisuuteen suuntautuvissa prosesseissa.

# 7 AR(1)-prosessi

AR(1)-prosessi määritellään seuraavasti

$$X_t - \mu = \phi(X_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{15}$$

jossa $\mu$ on prosessin odotusarvo, $\phi$ on autoregressiokerroin, ja $\varepsilon_t$ on normaalijakautunut virhetermi, joka on valkoista kohinaa. Stationaarisuuden varmistamiseksi oletetaan, että karakteristisen polynomin juuret sijaitsevat yksikköympyrän ulkopuolella, ks. Kohta 5.1.2, jolloin AR(1)-prosessin tapauksessa saadaan

$$1 - \phi z = 0 \iff z = \frac{1}{\phi}, \tag{16}$$

josta seuraa stationaarisuusehto

$$\left|\frac{1}{\phi}\right| > 1 \iff |\phi| < 1. \tag{17}$$

Tässä mallissa nykyinen havainto $X_t$ riippuu suoraan edellisestä arvosta $X_{t-1}$, mikä tarkoittaa, että perinteisen KRL:n riippumattomuusoletus ei päde. Keskeisen raja-arvolauseen (KRL) soveltaminen tällaisissa tapauksissa, kuten AR(1)-prosesseissa, on kuitenkin hyvin tärkeää, koska se mahdollistaa estimaattien normaalijakautuneisuuden suurilla otoksilla, vaikka havaintoarvot olisivat riippuvaisia keskenään, sekä etenkin luottamusvälien ja malliparametrien estimoinnissa, kuten $\phi$:n ja $\sigma^2$:n arvioinnissa. Tässä työssä keskitytään kuitenkin erityisesti mallin odotusarvoparametrin $\mu$ estimointiin. Vaikka yksittäiset havainnot eivät olisi normaalijakautuneita, niiden summat tai keskiarvot voivat asymptoottisesti noudattaa normaalijakaumaa, mikä on olennaista monille aikasarja-analyysin menetelmille.

Monissa sovelluksissa, kuten osaketuottojen ja korkojen mallintamisessa, AR(1)-prosesseja käytetään kuvaamaan ajallisesti riippuvaisia ilmiöitä. KRL mahdollistaa tilastollisten mallien käytön ennustamiseen ja riskien arviointiin. Monimutkaisemmissa malleissa, joissa käytetään AR(1)-prosessia osana regressiomallia, KRL auttaa varmistamaan, että jäännöstermit tai parametriestimaatit noudattavat normaalijakaumaa, mikä tekee hypoteesitestauksesta ja estimoinnista luotettavaa. Esimerkiksi, vaikka AR(1)-malli rikkoo riippumattomuusoletusta, KRL antaa perustan sille, että suuret otokset ja näiden keskiarvot voidaan silti käsitellä normaalijakaumaa approksimoivina. Tämä kuitenkin edellyttää erityisehtoja, joita käsitellään seuraavissa osioissa. Tämä tekee normaalijakauman ja sen johdannaisten, kuten $t$-testien ja $F$-testien, soveltamisen mahdolliseksi myös aikasarja-aineistoihin.

## 7.1 AR(1)-KRL

AR(1)-prosessissa (15) nykyinen arvo riippuu edellisestä arvosta, mikä rikkoo riippumattomuusoletusta. Vaikka riippumattomuusoletus puuttuu, AR(1)-prosessin keskiarvo saavuttaa ajan myötä asymptoottisen normaalisuuden, kun seuraavat ehdot täyttyvät (*ks. [15] s.1.*). Prosessin on oltava stationaarinen, mikä toteutuu silloin kun $|\phi| < 1$

ja virhetermien varianssi on vakio ja äärellinen. Stationaarisessa tapauksessa prosessin pitkän aikavälin odotusarvo ja varianssi ovat

$$E[X_t] = \mu, \quad \mathrm{Var}(X_t) = \frac{\sigma^2}{1-\phi^2} < \infty. \tag{18}$$

Tällöin AR(1)-keskeisen raja-arvolauseen mukaisesti prosessin summan normalisoitu keskiarvo konvergoi normaalijakaumaan

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} X_t - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{(1-\phi)^2}\right), \text{ kun } n \to \infty,$$

jossa $\mu$ on prosessin odotusarvo.

### 7.1.1 AR(1)-KRL todistus

Oletetaan, että AR(1)-malli on muotoa 15, jossa virhetermi $\varepsilon_t \sim$ i.i.d$(0, \sigma^2)$, jonka odotusarvo on 0 ja varianssi $\sigma^2$. Lisäksi $|\phi| < 1$, eli prosessi on stationaarinen. Tällöin otoskeskiarvo $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ noudattaa asymptoottista normaalijakaumaa

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{(1-\phi)^2}\right), \text{ kun } n \to \infty.$$

*Todistus.* Tämä todistus perustuu esitykseen lähteessä (*[15], s. 2–3*).
Määritellään

$$b_j = \sum_{k=j+1}^{\infty} \phi^k.$$

Tällöin

$$b_{j-1} - b_j = \sum_{k=j}^{\infty} \phi^k - \sum_{k=j+1}^{\infty} \phi^k = \phi^j.$$

Määritellään uusi lineaarinen aikasarja

$$Y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}.$$

Lineaarinen aikasarja $Y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}$ on hyvin määritelty, kun sillä on äärellinen varianssi. Koska $Y_t$ esiintyy myöhemmin osana $X_t$:n esitystä, on olennaista varmistaa, että kyseessä on stationaarinen prosessi. Tämä saavutetaan tarkistamalla, että virheiden painokertoimien neliösumma konvergoi. Tässä tapauksessa ehto täyttyy, koska sarja konvergoi, kun $|\phi| < 1$.

$$\sum_{j=0}^{\infty} b_j^2 = \sum_{j=0}^{\infty}\left(\sum_{k=j+1}^{\infty} \phi^k\right)^2 = \sum_{j=0}^{\infty}\left(\frac{\phi^{j+1}}{1-\phi}\right)^2 = \frac{\phi^2}{(1-\phi)^2}\sum_{j=0}^{\infty}\phi^{2j} = \frac{\phi^2}{(1-\phi)^2}\cdot\frac{1}{1-\phi^2} < \infty.$$

Koska $\sum_{j=0}^{\infty} b_j^2 < \infty$, lineaarinen prosessi $Y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}$ on määritelty ja stationaarinen,

$$\begin{aligned}
X_t &= \mu + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \\
&= \mu + \varepsilon_t + \sum_{j=1}^{\infty} \phi^j \varepsilon_{t-j} \\
&= \mu + \varepsilon_t + \sum_{j=1}^{\infty} (b_{j-1} - b_j) \varepsilon_{t-j} \\
&= \mu + \left( \sum_{j=0}^{\infty} \phi^j \right) \varepsilon_t - \left( \sum_{j=1}^{\infty} \phi^j \right) \varepsilon_t + \sum_{j=1}^{\infty} b_{j-1} \varepsilon_{t-j} - \sum_{j=1}^{\infty} b_j \varepsilon_{t-j} \\
&= \mu + \left( \sum_{j=0}^{\infty} \phi^j \right) \varepsilon_t - b_0 \varepsilon_t - \sum_{j=1}^{\infty} b_j \varepsilon_{t-j} + \sum_{j=1}^{\infty} b_{j-1} \varepsilon_{t-j} \\
&\text{(Olkoon} \quad k = j-1, j = k+1) \\
&= \mu + \left( \sum_{j=0}^{\infty} \phi^j \right) \varepsilon_t - \sum_{j=0}^{\infty} b_j \varepsilon_{t-j} + \sum_{k=0}^{\infty} b_k \varepsilon_{t-k-1} \\
&= \mu + \left( \sum_{j=0}^{\infty} \phi^j \right) \varepsilon_t - Y_t + Y_{t-1},
\end{aligned}$$

$$\text{joten} \quad X_t - \mu = \left( \sum_{j=0}^{\infty} \phi^j \right) \varepsilon_t - Y_t + Y_{t-1}.$$

Otetaan $\frac{1}{n} \sum_{t=1}^{n}$ molemmilta puolilta

$$\begin{aligned}
\frac{1}{n} \sum_{t=1}^{n} X_t - \mu \iff & \left( \sum_{j=0}^{\infty} \phi^j \right) \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t - \frac{1}{n} \sum_{t=1}^{n} Y_t + \frac{1}{n} \sum_{t=1}^{n} Y_{t-1} \\
\bar{X} - \mu \iff & \left( \frac{1}{1-\phi} \right) \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t - \frac{1}{n} (Y_n - Y_0) \\
\sqrt{n}(\bar{X} - \mu) \iff & \left( \frac{1}{1-\phi} \right) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t - \frac{1}{\sqrt{n}} (Y_n - Y_0).
\end{aligned}$$

Koska $\varepsilon_t \sim \text{i.i.d}(0, \sigma^2)$, niin KRL:n nojalla

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t = \sqrt{n}(\bar{\varepsilon}_n - 0) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

jossa $\bar{\varepsilon}_n = \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t$.
Lisäksi

$$\frac{1}{\sqrt{n}} Y_n \xrightarrow{p} 0, \quad \frac{1}{\sqrt{n}} Y_0 \xrightarrow{p} 0, \quad \text{kun} \quad n \to \infty.$$

Tästä seuraa, että AR(1)-keskeisen raja-arvolauseen mukaisesti prosessin summan normalisoitu keskiarvo

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

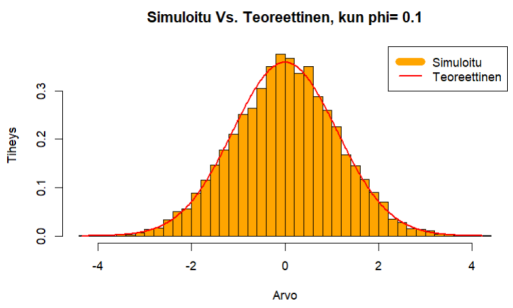
on asymptoottisesti normaalijakautunut

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{(1-\phi)^2}\right), \text{ kun } n \to \infty$$

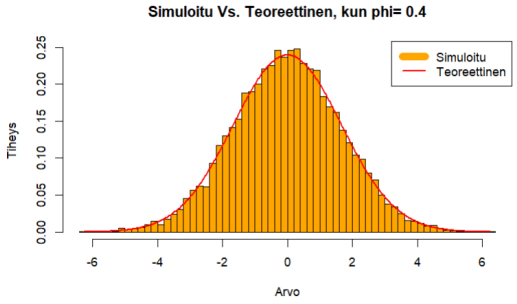
□

### 7.1.2 Esimerkki stationaarisuuden vaikutuksesta, AR(1)-KRL

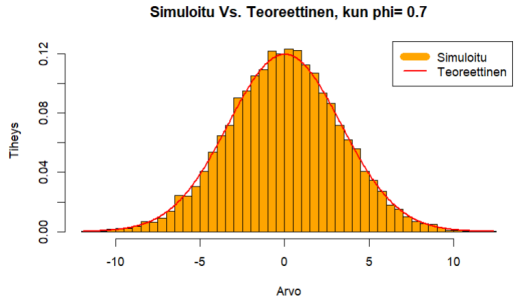
Alla on viisi kuvaajaa, jotka havainnollistavat AR(1)-prosessin keskiarvojen käyttäytymistä stationaarisuuden ($|\phi| < 1$) kasvaessa kohti epästationaarista $|\phi| \geq 1$ tilannetta. Simulaatioissa käytettiin seuraavia $|\phi|$:n arvoja: $(0.1, 0.4, 0.7, 0.9, 0.99)$. Kaikki simulaatiot toteutettiin otoskoossa $n = 100$ ja simulaatioiden määrä oli $m = 10,000$. Lisäksi oletettiin, että prosessin odotusarvo on $\mu = 0$ ja varianssi $\sigma^2 = 1$. Jokaisessa simulaatiossa generoitiin $n$-pituinen AR(1)-aikasarja, ja laskettiin sen keskiarvo $\bar{X}$. Keskiarvo standardoitiin $\sqrt{n}(\bar{X} - \mu)$, jossa $\mu$ on prosessin odotusarvo. Standardoiduista arvoista muodostettiin histogrammi (kuvaajissa oranssit pylväät).



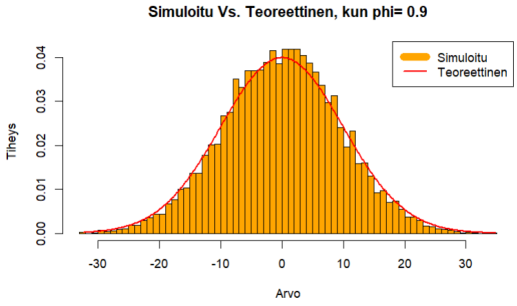

($|\phi| = 0.1$)




($|\phi| = 0.4$)

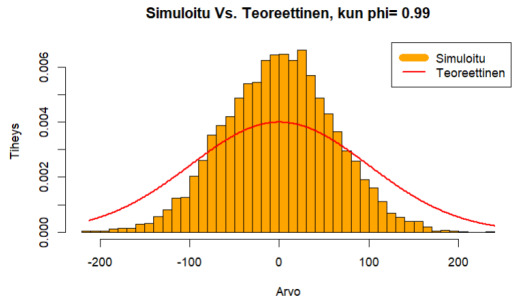



($|\phi| = 0.7$)




($|\phi| = 0.9$)

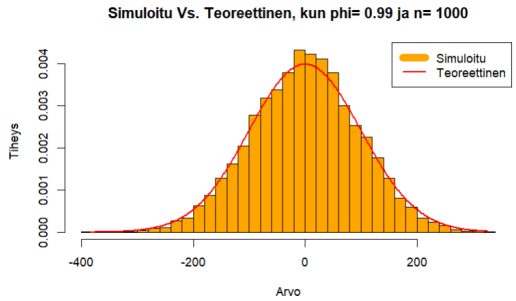
Kuva 13: Stationaariset tapaukset, AR(1)-KRL: havainnollistaminen




$(|\phi| = 0.99)$

Kuva 14: Stationaariset tapaukset, AR(1)-KRL: havainnollistaminen

Stationaarisissa tapauksissa (kuva 13), joissa $|\phi| < 1$ prosessi noudattaa AR(1)-keskeisen raja-arvolauseen ennustetta jo hyvin pienellä otoskoolla $n = 100$. Histogrammit osoittavat, että normaaliapproksimaatio seuraa teoreettista normaalijakaumaa ja epästationaarisuuden vaikutus alkaa olla havaittavissa vasta kun $|\phi| = 0.99$, kuva 14. Mikäli otoskokoa kasvatetaan esimerkiksi tuhanteen, kuten kuva 15 osoittaa, niin normaaliapproksimaatio seuraa teoreettista normaalijakaumaa erittäin tarkasti vielä $|\phi|$:n ollessa 0.99. Toisin sanoen, mitä isompi $\phi$, sitä pidempi sarja tarvitaan, jotta approksimaatio olisi tarpeeksi tarkka.




Kuva 15: AR(1)-KRL, kun $n = 1000$

Prosessin varianssi määritellään kuvissa 13 - 15 aikaisemmin esitetyllä varianssin kaavalla (18), jossa autoregressiivisen kertoimen $|\phi|$ arvot vaihtelevat välillä $0.1 - 0.99$ ja varianssin $\frac{\sigma^2}{(1-\phi)^2}$ arvot vastaavasti välillä $1.01 - 50.25$

Epästationaarisessa tapauksessa $|\phi| \geq 1$ prosessin varianssi kasvaa rajatta ajan myötä (*[13] s. 256-265*), kun

$$\mathrm{Var}(X_t) \to \infty.$$

Tämä ilmiö näkyy mallin varianssin kaavassa siten, että varianssi (18) ei ole mielekäs, kun $1 - \phi^2 \leq 0$. Erityisesti silloin, kun $\phi = 1$, nimittäjä muuttuu nollaksi, mikä tekee varianssin määrittelemättömäksi, ja mikäli $\phi > 1$, jolloin nimittäjä on negatiivinen, mikä on seurausta epästationaarisen prosessin eksponentiaalisesta kasvusta.

# 8 AR(2)-prosessi

AR(2)-prosessi voidaan esittää seuraavasti

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

jossa $X_t$ on aikasarjan arvo hetkellä $t$ ja $\mu$ on prosessin odotusarvo. Autoregressiiviset kertoimet $\phi_1, \phi_2$ kerrotaan viivästetyillä havainnoilla $X_{t-1}, X_{t-2}$ ja $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ on normaalisti jakautunut virhetermi, joka on valkoista kohinaa. Kuten kohdassa 5.1.2 todettiin, AR($p$)-prosessin stationaarisuus perustuu karakteristisen polynomien juuriin (9), joiden pitää sijaita yksikköympyrän ulkopuolella $|z| > 1$. Lähteeseen *[34]* nojaten, AR(2)-prosessin tapauksessa tämä voidaan esittää seuraavan yhtälön muodossa

$$1 - \phi_1 z - \phi_2 z^2 = 0. \tag{19}$$

Tämän yhtälön juuret voidaan ratkaista toisen asteen yhtälön ratkaisukaavalla

$$\left| \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \right| > 1.$$

Mikäli juuret sijaitsevat yksikköympyrän ulkopuolella, seuraa tästä tunnetut kertoimiin perutuvat ehdot

$$\begin{aligned} \phi_1 + \phi_2 &< 1, \\ \phi_2 - \phi_1 &< 1, \\ |\phi_2| &< 1 \end{aligned} \tag{20}$$

*[34]*. Tämä malli on erityisen hyödyllinen aikasarja-analyysissä, kun havaintojen väliset vaikutukset ja viivästykset tekevät yksinkertaisten kausaaliyhteyksien määrittämisestä vaikeaa. AR(2)-malli kykenee kuvaamaan prosessin dynamiikkaa tarkemmin kuin yksinkertaisempi AR(1)-malli, erityisesti silloin, kun aikasarjassa esiintyy monivaiheisia riippuvuuksia ja viiveiden vaikutukset ovat merkittäviä. AR(2)-mallissa stationaarisen varianssin laskeminen on kuitenkin hieman mutkikkaampaa kuin esimerkiksi AR(1)-mallissa, ja se perustuu autokovarianssien (7) ja autokorrelaatioiden (8) viiveoperaattoreihin, jotka lasketaan Yule-Walker-menetelmällä. Seuraavassa on esitetty AR(2)-prosessin varianssin johtamisen todistus, joka perustuu tähän menetelmään. Todistus perustuu lähteeseen *[2]*.

**Lause 1.** *AR(2)-mallin varianssi on* $Var(X_t) = \frac{\sigma^2(1-\phi_2)}{(1-\phi_1^2-\phi_2^2)(1-\phi_2)-2\phi_1^2\phi_2}$.

*Todistus.*

$$\begin{aligned} Var(X_t) &= Var(\phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t) \\ &= \phi_1^2 Var(X_{t-1}) + \phi_2^2 Var(X_{t-2}) + Var(\varepsilon_t) + 2\phi_1\phi_2 Cov(X_{t-1}, X_{t-2}) \\ &= \phi_1^2 Var(X_t) + \phi_2^2 Var(X_t) + \sigma^2 + 2\phi_1\phi_2 Cov(X_{t-1}, X_{t-2}) \end{aligned}$$

Tällöin stationaarisuuden perusteella

$$Var(X_t) = \phi_1^2 Var(X_t) + \phi_2^2 Var(X_t) + \sigma^2 + 2\phi_1\phi_2\rho_1 Var(X_t),$$

saadaan

$$Var(X_t) = \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\rho_1}.$$

Kuitenkin ominaisuuden (8) mukaan $\rho_1 = \frac{\phi_1}{1-\phi_2}$, jolloin lopulliseksi varianssiksi AR(2)-prosessille muodostuu

$$Var(X_t) = \gamma_0 = \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1\phi_2\frac{\phi_1}{1-\phi_2}} = \frac{\sigma^2(1-\phi_2)}{(1-\phi_1^2-\phi_2^2)(1-\phi_2) - 2\phi_1^2\phi_2} \tag{21}$$

□

## 8.1 AR(2)-KRL

Samoin kuten AR(1)-prosessissa, voidaan AR(2)-prosessin keskiarvoille $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ käyttää asymptoottista normaalijakaumaa, kun $n \to \infty$. Tämä tarkoittaa, että otoskoon kasvaessa prosessin keskiarvojen jakauma lähestyy normaalijakaumaa, vaikka alkuperäinen jakauma ei olisikaan normaalisti jakautunut. Toisin sanoen, KRL:n nojalla, vaikka yksittäiset havainnot $X_t$ eivät olisi normaalisti jakautuneita, niiden keskiarvot kuitenkin noudattavat asymptoottisesti normaalijakaumaa, mikä mahdollistaa normaalijakautuneiden estimointimenetelmien käytön suurilla otosmäärillä. Vaikka AR(2)-prosessin havainnot eivät ole toisistaan riippumattomia, otoskeskiarvo voi silti lähestyä normaalijakaumaa, kunhan sillä on vakio ja äärellinen varianssi. Stationaarisuuden ehdot voi tarkistaa esimerkiksi karakteristisen yhtälön juurten avulla (20).

Kaava (21) kuvaa tilannetta yksittäisen havainnon $X_t$ varianssista, joka saadaan Yule-Walker-yhtälöiden avulla. AR(2)-prosessin tapauksessa keskiarvon varianssi ei kuitenkaan ole yksinkertaisesti $Var(X_t)/n$, kuten riippumattomien havaintojen tapauksessa. Tämä johtuu siitä, että prosessin havainnot ovat autokorreloituneita, eli niiden välillä on ajallista riippuvuutta. Tämän vuoksi keskiarvon varianssiin vaikuttavat myös kaikkien aikaviiveiden yhteisvaikutukset. Nämä vaikutukset voidaan huomioida MA($\infty$) -esityksen kautta, jossa AR(2)-prosessi esitetään äärettömänä summana.

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i},$$

jossa painotus tapahtuu $\psi_i$-kertoimilla, kun $i = 0, 1, 2, ...$, jolloin prosessi noudattaa MA($\infty$)-prosessia *[3]*.

Lähteen *[11]* lauseen 7.1.2 mukaan keskiarvon asymptoottinen varianssi saadaan ottamalla näiden MA-kertoimien summan neliö ja kertomalla se $\sigma^2$:lla, jolloin asymptoottiseksi varianssiksi saadaan

$$Var_{asymp} = \sigma^2 \cdot \left(\sum_{t=0}^{\infty} \psi_i\right)^2 < \infty. \tag{22}$$

Tarkempi todistus esitetään lähteen *[11]* osiossa 7.3.

Seuraavassa lähteenä on käytetty Florian Kölblin diplomityötä *Aggregation of AR(2) Processes [3]*, jossa osoitetaan, että summat $\sum_{i=0}^{\infty}\psi_i$, $\sum_{i=0}^{\infty}\psi_i^2$, $\sum_{i=0}^{\infty}\psi_i\psi_{i+h}$ ovat äärellisiä, jos karakteristisen yhtälön $\lambda^2 - \phi_1\lambda - \phi_2 = 0$ juuret $\lambda_1$ ja $\lambda_2$, jossa $\phi_1 = \lambda_1 + \lambda_2$

ja $\phi_2 = -\lambda_1\lambda_2$ ovat yksikköympyrän sisäpuolella. Tällöin

$$\psi_i = \frac{\lambda_1^{i+1} - \lambda_2^{i+1}}{\lambda_1 - \lambda_2},$$

ja saadaan, että

$$\begin{aligned}\sum_{i=0}^{\infty} \psi_i &= \sum_{i=0}^{\infty} \frac{\lambda_1^{i+1} - \lambda_2^{i+1}}{\lambda_1 - \lambda_2} = \frac{1}{\lambda_1 - \lambda_2} \sum_{i=1}^{\infty} (\lambda_1^i - \lambda_2^i) \\ &= \frac{1}{\lambda_1 - \lambda_2} \sum_{i=0}^{\infty} (\lambda_1^i - \lambda_2^i) = \frac{1}{\lambda_1 - \lambda_2} \left( \frac{1}{1 - \lambda_1} - \frac{1}{1 - \lambda_2} \right) \\ &= \frac{1}{(1 - \lambda_1)(1 - \lambda_2)} = \frac{1}{(1 - \phi_1 - \phi_2)} \end{aligned} \tag{23}$$

*[3]*. Sijoittamalla kaava (23) kaavan (22) sulkeiden sisäpuolelle, saadaan AR(2)-prosessin asymptoottinen varianssi, joka saa muodon

$$Var_{asymp} = \frac{\sigma^2}{(1 - \phi_1 - \phi_2)^2}.$$

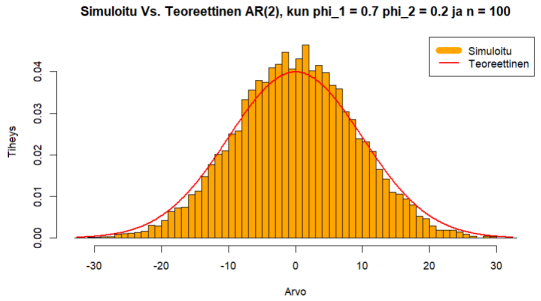
Tällöin AR(2)-keskeisen raja-arvolauseen mukaisesti prosessin summan normalisoitu keskiarvo konvergoi normaalijakaumaan seuraavasti

$$\sqrt{n} \left( \frac{1}{n} \sum_{t=1}^{n} X_t - \mu \right) \xrightarrow{d} \mathcal{N} \left( 0, \frac{\sigma^2}{(1 - \phi_1 - \phi_2)^2} \right), \text{ kun } n \to \infty.$$

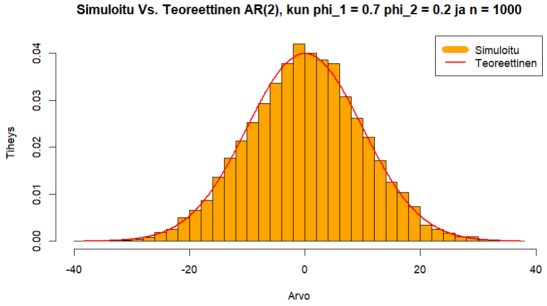
Asymptoottisen normaalijakauman varianssista voidaan tehdä huomioita liittyen autoregressiivisten kertoimien $\phi_1$ ja $\phi_2$ merkitykseen. Varianssin kaavasta nähdään, että molemmat kertoimet vaikuttavat siihen samalla tavalla. Esimerkiksi, jos $\phi_1$ kasvaa 0.1:llä, varianssi muuttuu täsmälleen saman verran kuin jos $\phi_2$ kasvaisi 0.1:llä. Tämä tarkoittaa, että molemmat kertoimet ovat yhtä kriittisessä asemassa estimoinnin tarkkuuden kannalta.

### 8.1.1 Esimerkki stationaarisuuden vaikutuksesta, AR(2)-KRL

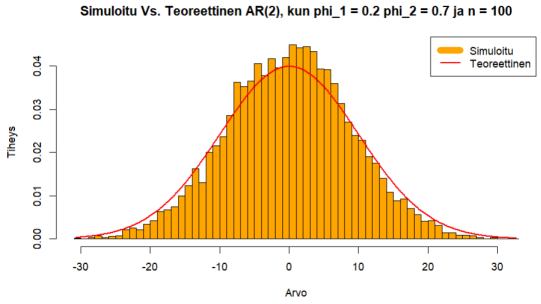
AR(2)-KRL:n käyttäytyminen riippuu kertoimista $\phi_1$ ja $\phi_2$. Alla on esimerkkinä kaksi simulaatiota, joissa tarkastellaan prosessin käyttäytymistä eri kertoimien arvoilla niin, että niiden vaikutusta voidaan vertailla rinnakkain. Käytännössä tämä tarkoittaa, että toisessa tapauksessa ensimmäinen viive $\phi_1$ on suurempi ja toinen viive $\phi_2$ pienempi, kun taas toisessa tapauksessa roolit on vaihdettu päinvastaisiksi. Molemmissa simulaatioissa säilytetään ehto $|\phi_1 + \phi_2| < 1$, jotta prosessi pysyy stationaarisena. Lisäksi on luotu kaksi kuvaajaa, joissa $n = 50$.



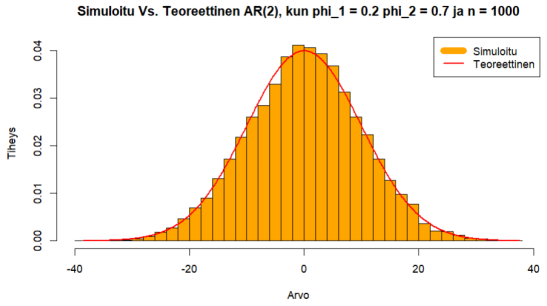

Simulaatio 1, kun $n = 100$




Simulaatio 1, kun $n = 1000$



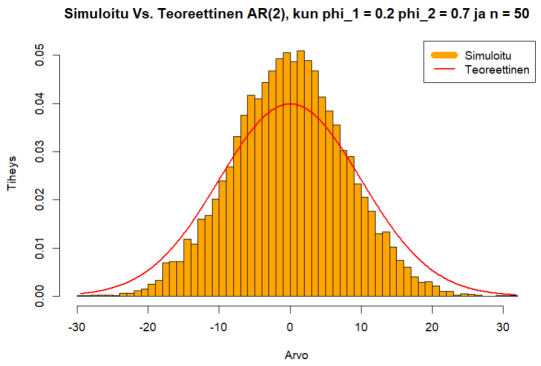

Simulaatio 2, kun $n = 100$



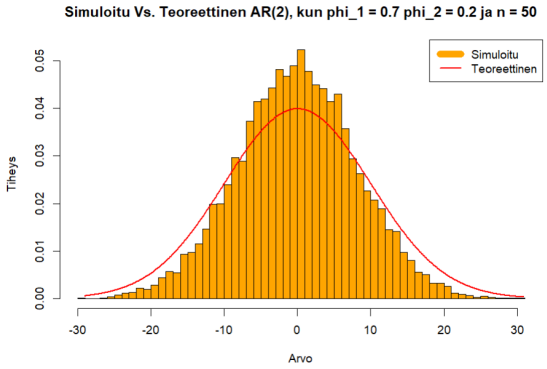

Simulaatio 2, kun $n =1000$

Kuva 16: Stationaariset tapaukset AR(2)-KRL

Ensimmäisessä simulaatiossa asetettiin $\phi_1 = 0.7$ ja $\phi_2 = 0.2$, jolloin stationaarisuusehto $|\phi_1 + \phi_2| < 1$ täyttyy. Toisessa simulaatiossa parametrit vaihdettiin päinvastoin, eli $\phi_1 = 0.2$ ja $\phi_2 = 0.7$, jolloin ehto säilyy edelleen voimassa. Molemmat simulaatiot toteutettiin kahdella eri otoskoolla, $n = 100$ ja $n = 1000$, ja simulaatioiden määrä oli $m = 10,000$. Lisäksi oletettiin, että prosessin odotusarvo on $E(X_t) = 0$ ja varianssi $\sigma^2 = 1$. Kuvassa 16 huomaa, että AR(2)-KRL:n normaaliapproksimaatio seuraa teoreettista normaalijakaumaa, kun $|\phi_1 + \phi_2| < 1$.




Simulaatio 1, kun $n = 50$




Simulaatio 2, kun $n = 50$

Kuva 17: Stationaariset tapaukset AR(2)-KRL, kun $n = 50$

Kuvassa 17 otoskoko on $n = 50$. Tässä käytettiin samaa menetelmää ja samoja $\phi$:n arvoja kuin kuvassa 16. Myös tässä AR(2)-keskeinen raja-arvolauseen normaaliapproksimaatio seuraa teoreettista normaalijakaumaa suhteellisen hyvin, kun $|\phi_1 + \phi_2| < 1$ vaikka otoskoko on hyvin pieni.

### 8.1.2 Johtopäätös

AR($p$)-prosessiin voidaan soveltaa keskeistä raja-arvolausetta ainoastaan silloin, kun prosessi on stationaarinen ja sen varianssi on vakio ja äärellinen. Näissä olosuhteissa pitkän aikavälin ominaisuudet, kuten keskiarvot ja varianssit, pysyvät vakaina, ja normalisoitu summa konvergoi normaalijakaumaan. Tämä mahdollistaa AR($p$)-prosessin asymptoottisen analyysin myös käytännön sovelluksissa.

Täysin epästationaarinen tapaus voitaisiin havainnollistaa tarkastelemalla, miten prosessin dynamiikka muuttuu, kun valitaan esimerkiksi $|\phi_1 + \phi_2| = 1.4 > 1$. Tällöin ehto $|\phi_1 + \phi_2| < 1$ ei enää pidä paikkaansa, mikä tarkoittaa, että prosessi ei ole enää stationaarinen. Tällaisessa tilanteessa prosessin varianssi kasvaa rajatta ajan myötä, mikä johtaa epävakaaseen käyttäytymiseen, esimerkiksi eksponentiaaliseen kasvuun.

# 9 MA(2)-prosessi

Tässä osiossa keskitytään MA(2)-KRL:ään, sillä MA(1)-KRL:n yksinkertainen autokorrelaatiokäyttäytyminen tulee esille jo MA(2)-mallin yhteydessä, joten MA(1)-mallia ei käsitellä erikseen.

MA(2)-prosessin määritelmä on seuraava

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Tässä mallissa nykyinen havainto $X_t$ riippuu suoraan korkeintaan kahdesta aikaisemmasta virhetermistä $\varepsilon_{t-1}$ ja $\varepsilon_{t-2}$, joten riippumattomuusoletus ei päde. Peräkkäiset havainnot ovat korreloituneita korkeintaan kahden viiveen verran.

MA(2)-prosessin kääntyvyysehdot puolestaan perustuvat aikaisemmin osiossa 5.2.2 käsiteltyihin MA($q$)-prosessin polynomeihin 13, jossa kaikkien juurten tulee sijaita yksikköympyrän ulkopuolella. MA(2)-prosessin tapauksessa tämä esitetään seuraavan yhtälön muodossa

$$1 - \theta_1 z - \theta_2 z^2 = 0.$$

Mikäli juuret sijaitsevat yksikköympyrän ulkopuolella $|z| > 1$, niin kääntyvyys voidaan varmistaa seuraavien epäyhtälöiden muodossa

$$|\theta_2| < 1, \quad \theta_1 + \theta_2 < 1, \quad \theta_2 - \theta_1 < 1, \tag{24}$$

Näiden ehtojen täyttyessä MA(2)-prosessi on kääntyvä. Mikäli kääntyvyysehdot eivät täyty, prosessi säilyy kyllä stationaarisena ja MA(2)-KRL:n mukainen asymptoottinen normaalisuus on periaatteessa voimassa, mutta mallin tulkinta ja parametriarvioiden käyttäytyminen voivat muuttua hankaliksi tai epäintuitiivisiksi. Täten kääntyvyys ei siis ole KRL:n muodollinen ehto. Kääntyvyysehdot kuitenkin huomioidaan lähes poikkeuksetta, kun MA($q$)-prosesseja käytetään tilastollisessa mallintamisessa, joten myös tässä työssä pyritään huomioimaan tämä seikka.

Koska kyseessä on MA(2)-prosessi, saadaan niiden autokovarianssit ja autokorrelaatiot soveltamalla aikaisemmin osiossa 5.2.1 käsitellyn MA($q$)-prosessin kaavoja 11 ja 12 seuraavasti

$$\begin{aligned}
\gamma_0 &= \sigma^2(1 + \theta_1^2 + \theta_2^2), \\
\gamma_1 &= \sigma^2(\theta_1 + \theta_1\theta_2), \\
\gamma_2 &= \sigma^2\theta_2, \\
\gamma_k &= 0, \quad |k| > 2,
\end{aligned} \tag{25}$$

ja autokorrelaatiot puolestaan seuraavasti

$$\begin{aligned}
\rho_0 &= 1, \\
\rho_1 &= \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \\
\rho_2 &= \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}, \\
\rho_k &= 0, \quad \text{kun } |k| > 2
\end{aligned}$$

*[28]*. Tästä voidaan havaita, että MA(2)-prosessin autokorrelaatiofunktio saa nollaarvon kaikilla viiveillä, jotka ylittävät kaksi.

## 9.1 MA(2)-KRL

Vaikka MA(2)-prosessin havainnot $X_t$ ovat riippuvaisia kahdesta aiemmasta virhetermistä, otoskeskiarvojen

$$\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$$

jakauma voi suurilla otoksilla lähestyä normaalijakaumaa MA(2)-KRL:n mukaisesti. Tällöin otoskeskiarvon asymptoottista varianssia voidaan arvioida summana autokovariansseista seuraavasti

$$Var_{Asymp} = \gamma_0 + 2\gamma_1 + 2\gamma_2 \tag{26}$$

*[11]*. Tämän summan voi laskea suljetussa muodossa

$$Var_{Asymp} = \sigma^2\left[1 + \theta_1^2 + \theta_2^2 + 2\theta_1 + 2\theta_1\theta_2 + 2\theta_2\right] = \sigma^2(1 + \theta_1 + \theta_2)^2.$$

Vaikka tässä ei esitetä täyttä todistusta KRL:n soveltuvuudesta MA(2)-prosessille, voidaan osoittaa (*Ks. Lähde [11] Esim. 6.4.3 ja 6.4.4. S.214*), että prosessin otoskeskiarvo on asymptoottisesti normaalijakautunut. Tälle varianssille pätee

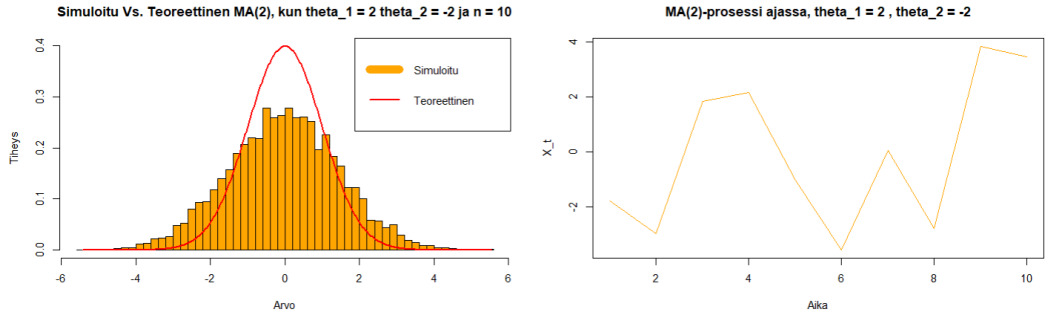
$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} X_t - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(1 + \theta_1 + \theta_2)^2\right), \text{ kun } n \to \infty.$$

## 9.2 Esimerkki stationaarisuuden vaikutuksesta, MA(2)-KRL

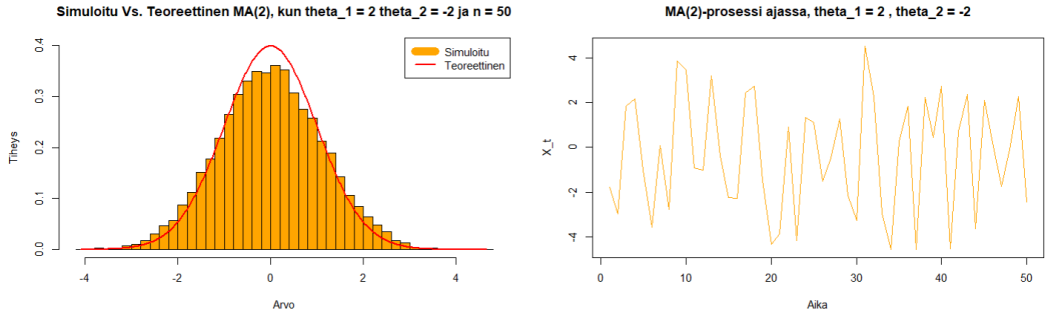
Kuten aikaisemmin todettiin, MA($q$)-prosessi on aina stationaarinen, joten myöskään MA(2)-KRL:n tapauksessa kääntyvyysehdolla ei ole suurta vaikutusta ja se täyttää KRL:n vaatimukset teoreettisesti. Parametrien suuruudella on kuitenkin käytännössä merkittävä vaikutus siihen, kuinka nopeasti keskiarvon jakauma lähestyy normaalijakaumaa. Pienillä ja kohtuullisilla $\theta$:n arvoilla keskiarvon jakauma on simulaatioissa usein hyvin normaalin näköinen jo pienillä otoksilla, kuten $n = 10$ ja $n = 100$. Tämä johtuu siitä, että $X_t$ on lineaarikombinaatio normaalisti jakautuneista kohinatermeistä ja keskiarvon vaihtelu tasoittuu nopeasti.

Jos kuitenkin parametrien suuruudet kasvavat huomattavasti, niin tilanne muuttuu oleellisesti, koska suuret parametrit kasvattavat prosessin autokovarianssia ja siten yksittäisten havaintojen $X_t$ vaihtelu voi olla hyvin suurta. Tällöin otoskoon tulee olla suurempi, esimerkiksi $n = 1000$, sillä yksittäiset suuria poikkeamia sisältävät havainnot voivat vaikuttaa otoskeskiarvoon merkittävästi, jolloin jakauman suppeneminen normaalimuotoon tapahtuu hitaammin.

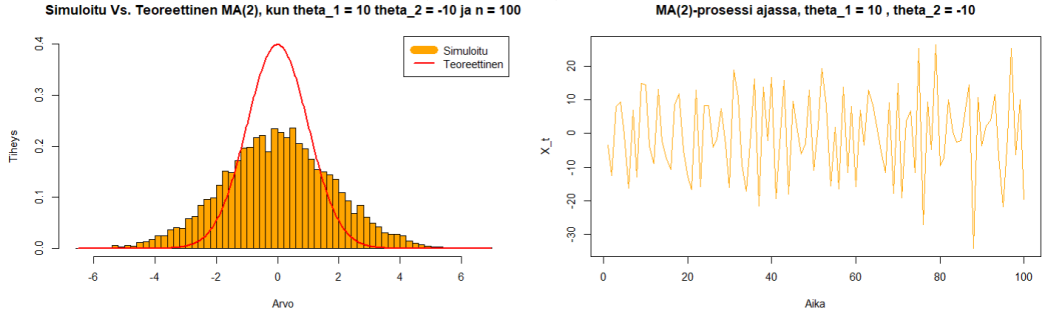
Tällainen lienee kuitenkin harvinaista reaalimaailmassa, koska tällöin menneisyyden vaikutus nykyhetkeen olisi poikkeuksellisen suurta ja data vaatisi tarkempaa tutkimista. Käytännössä kääntyvyysehdot rajoittavat $\theta$:n arvoja jo itsessään. Vaikka MA($q$)-KRL ei sitä vaadi, tilastollinen mallinnus käytännössä edellyttää kääntyvyyttä, jolloin suuria $\theta$:n arvoja ei edes hyväksyttäisi malliin.



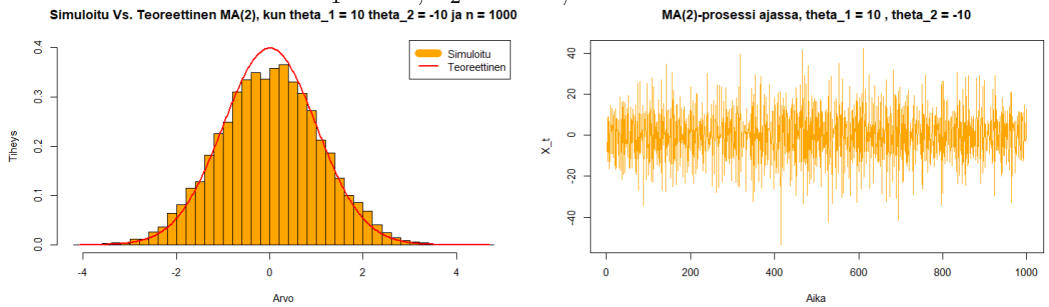

$\theta_1 = 2$, $\theta_2 = -2$, kun $n = 10$




$\theta_1 = 2$, $\theta_2 = -2$, kun $n = 50$




$\theta_1 = 10$, $\theta_2 = -10$, kun $n = 100$




$\theta_1 = 10$, $\theta_2 = -10$, kun $n =$1000

Kuva 18: MA(2)-KRL eri liukuvan keskiarvon parametreillä ja niitä vastaavat aikasarjakuvaajat

Kuvassa 18 on esitetty tilanteita eri liukuvien keskiarvojen parametreillä. Kummassakaan tapauksessa kääntyvyysehto ei toteudu. Kuvista voi havaita, että MA(2)-KRL toimii hyvin jo pienillä otoksilla, kunhan liukuvan keskiarvon parametrit ovat kohtuullisia $\theta_1 = 2$ ja $\theta_2 = -2$. Epärealistisessa tapauksessa $\theta_1 = 10$ ja $\theta_2 = -10$ MA(2)-KRL toimii myös hyvin, kunhan $n$ on tarpeeksi suuri. Myöhemmin simulaatio-osiossa tarkas-

tellaan ensimmäisen tapauksen osalta peittotodennäköisyyksiä. Nämä on havainnollistettu kuvassa 22 ja tarkemmat arvot näkee taulukosta 2. Näistä käy ilmi, että arvoilla $\theta_1 = 2$ ja $\theta_2 = -2$ jakauma konvergoi hitaammin kohti nimellistasoa pienillä otoksilla, mutta saavuttaa halutun tason lopulta otoskoon kasvaessa.

# 10 ARMA(1,1)-prosessi

ARMA(1,1)-prosessi määritellään seuraavasti

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1},\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

jonka odotusarvo on muotoa

$$E[X_t] = \mu.$$

Kun oletuksena on, että $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, niin ARMA(1,1)-prosessin varianssi on muotoa

$$Var(X_t) = Var(X_t - \mu) = \left(1 + \frac{(\phi_1 + \theta_1)^2}{(1 - \phi_1^2)}\right)\sigma^2$$

*[4]*. Mikäli mallin odotusarvo on $\mu = 0$, niin malli voidaan esittää muodossa

$$X_t = \phi_1 X_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Viitaten osioon 5.3.2, jossa käsiteltiin ARMA$(p, q)$-mallin karakteristisia juuria, ARMA-(1,1)-prosessissa tulee huomioida erikseen AR(1)- ja MA(1)-osan ehdot. AR(1)-osan juurten (16) tulee sijoittua yksikköympyrän ulkopuolelle, mikä takaa prosessin stationaarisuuden (17). MA(1)-osan juurten puolestaan tulee sijoittua yksikköympyrän ulkopuolelle kääntyvyyden varmistamiseksi, mikä johtaa määritelmän (14) mukaiseen epäyhtälöön.
Seuraavaksi esitetään ARMA(1,1):n autokovarianssit

$$\gamma_0 = \sigma^2\frac{(1 + 2\theta_1\phi_1 + \theta_1^2)}{(1 - \phi_1^2)},$$
$$\gamma_1 = \sigma^2\frac{(1 + \theta_1\phi_1)(\phi_1 + \theta_1)}{(1 - \phi_1^2)},$$

jossa $\gamma(k) = \phi_1^{k-1}\gamma(1)$, $k \geq 2$ *[34]*.
Autokorrelaatiot ARMA(1,1)-prosessille puolestaan ovat seuraavat

$$\rho_k = \begin{cases} 1, & k = 0 \\ \frac{(1+\theta_1\phi_1)(\phi_1+\theta_1)}{1+2\theta_1\phi_1+\theta_1^2}, & k = 1 \\ \phi_1\rho_{k-1}, & k \geq 2 \end{cases}$$

*[1]*. Vaikka ARMA(1,1)-prosessi on riippuvainen aikaisemmista havainnoista ja virhetermeistä, niin sen keskiarvojen $\bar{X}$ jakauma voi lähestyä normaalijakaumaa, kun $n \to \infty$. Tämä edellyttää, että perusoletukset, kuten stationaarisuus ovat voimassa. Asymptoottinen varianssi voidaan määritellä, kuten edellisessä osiossa laskettaessa MA(2)-KRL:n asymptoottista varianssia (26). Tällöin, kun lisäksi otetaan huomioon ääretön summa, saadaan

$$Var_{Asymp} = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k = \sigma^2\left(\frac{1 + \theta_1}{1 - \phi_1}\right)^2, \tag{27}$$

jossa ARMA(1,1)-prosessille $\gamma_k = \phi_1^{k-1}\gamma_1$, $k \geq 2$ *[34]*. Asymptoottinen varianssi $Var_{Asymp}$ saadaan alla olevan todistuksen mukaisesti.

## 10.1 ARMA(1,1)-KRL ja asymptoottisen varianssin todistus

**Lause 2.** *ARMA(1,1)-KRL:n asymptoottinen varianssi on* $Var_{Asymp} = \sigma^2 \left(\frac{1+\theta_1}{1-\phi_1}\right)^2$.

*Todistus.* Seuraavassa todistetaan ARMA(1,1) asymptoottinen varianssi, josta käy ilmi yllä olevan kaavan (27) alkuperä. Aloitetaan autokovariansseista $\gamma_0$ ja $\gamma_1$ jotka ovat muotoa

$$\gamma_0 = \sigma^2 \frac{(1 + 2\theta_1\phi_1 + \theta_1^2)}{(1 - \phi_1^2)} \quad \text{ja} \quad \gamma_1 = \sigma^2 \frac{(1 + \theta_1\phi_1)(\phi_1 + \theta_1)}{(1 - \phi_1^2)}.$$

Kun $k \geq 2$ autokovarianssien välinen suhde on $\gamma_k = \phi_1\gamma_{k-1}$, jolloin se on eksponentiaalisesti vähenevä $\gamma_k = \phi_1^{k-1}\gamma_1$. Asymptoottinen varianssi ARMA(1,1):lle on tällöin

$$Var_{\text{Asymp}} = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k = \gamma_0 + 2[\gamma_1 + \gamma_2 + \gamma_3 + \cdots].$$

Korvataan $\gamma_k = \phi_1^{k-1}\gamma_1$, jolloin

$$Var_{\text{Asymp}} = \gamma_0 + 2\sum_{k=1}^{\infty} \phi_1^{k-1}\gamma_1 = \gamma_0 + 2\gamma_1 \sum_{k=1}^{\infty} \phi_1^{k-1}.$$

Geometrinen sarja alkaa $k = 1$, joten

$$\sum_{k=1}^{\infty} \phi_1^{k-1} = \frac{1}{1 - \phi_1}.$$

Tästä saadaan

$$\begin{aligned} Var_{Asymp} &= \gamma_0 + \frac{2}{1 - \phi_1} \cdot \gamma_1 \\ &= \sigma^2 \frac{(1 + 2\theta_1\phi_1 + \theta_1^2)}{(1 - \phi_1^2)} + \frac{2}{1 - \phi_1} \left( \sigma^2 \frac{(1 + \theta_1\phi_1)(\phi_1 + \theta_1)}{(1 - \phi_1^2)} \right). \end{aligned}$$

Yhdistetään nimittäjät $(1 - \phi_1^2)(1 - \phi_1) = (1 - \phi_1)^2(1 + \phi_1)$, jolloin

$$Var_{Asymp} = \frac{\sigma^2}{(1 - \phi_1)^2(1 + \phi_1)}[(1 + \theta_1^2 + 2\phi_1\theta_1)(1 - \phi_1) + 2(\phi_1 + \theta_1)(1 + \phi_1\theta_1)].$$

Hakasulkeiden sisällä olevien termien laventaminen tuottaa

$$(1 + \theta_1^2 + 2\phi_1\theta_1)(1 - \phi_1) = 1 + \theta_1^2 + 2\phi_1\theta_1 - \phi_1 - \phi_1\theta_1^2 - 2\phi_1^2\theta_1$$

ja

$$2(\phi_1 + \theta_1)(1 + \phi_1\theta_1) = 2(\phi_1 + \theta_1 + \phi_1^2\theta_1 + \phi_1\theta_1^2).$$

Termien yhdistäminen ja lausekkeen järjestely tuottaa

$$1 + \theta_1^2 + 2\theta_1 + \phi_1 + 2\phi_1\theta_1 + \phi_1\theta_1^2 = (1 + 2\theta_1 + \theta_1^2) + \phi_1(1 + 2\theta_1 + \theta_1^2).$$

Koska $(1 + 2\theta_1 + \theta_1^2) = (1 + \theta_1)^2$, niin lopulliseksi muodoksi saadaan $(1 + \theta_1)^2(1 + \phi_1)$. Pääkaavaan takaisin sijoittamalla saadaan

$$\begin{aligned} Var_{Asymp} &= \frac{\sigma^2}{(1-\phi_1)^2(1+\phi_1)} \cdot (1+\theta_1)^2(1+\phi_1) \\ &= \frac{\sigma^2(1+\theta_1)^2(1+\phi_1)}{(1-\phi_1)^2(1+\phi_1)} \\ &= \sigma^2 \left( \frac{(1+\theta_1)^2}{(1-\phi_1)^2} \right) = \sigma^2 \left( \frac{1+\theta_1}{1-\phi_1} \right)^2. \end{aligned}$$

Kun $n \to \infty$, niin prosessin otoskeskiarvot lähestyvät asymptoottisesti normaalijakaumaa, jolloin KRL:n mukaan pätee

$$\sqrt{n} \left( \frac{1}{n} \sum_{t=1}^{n} X_t - \mu \right) \xrightarrow{d} \mathcal{N} \left( 0, \sigma^2 \left( \frac{1+\theta_1}{1-\phi_1} \right)^2 \right), \text{ kun } n \to \infty.$$

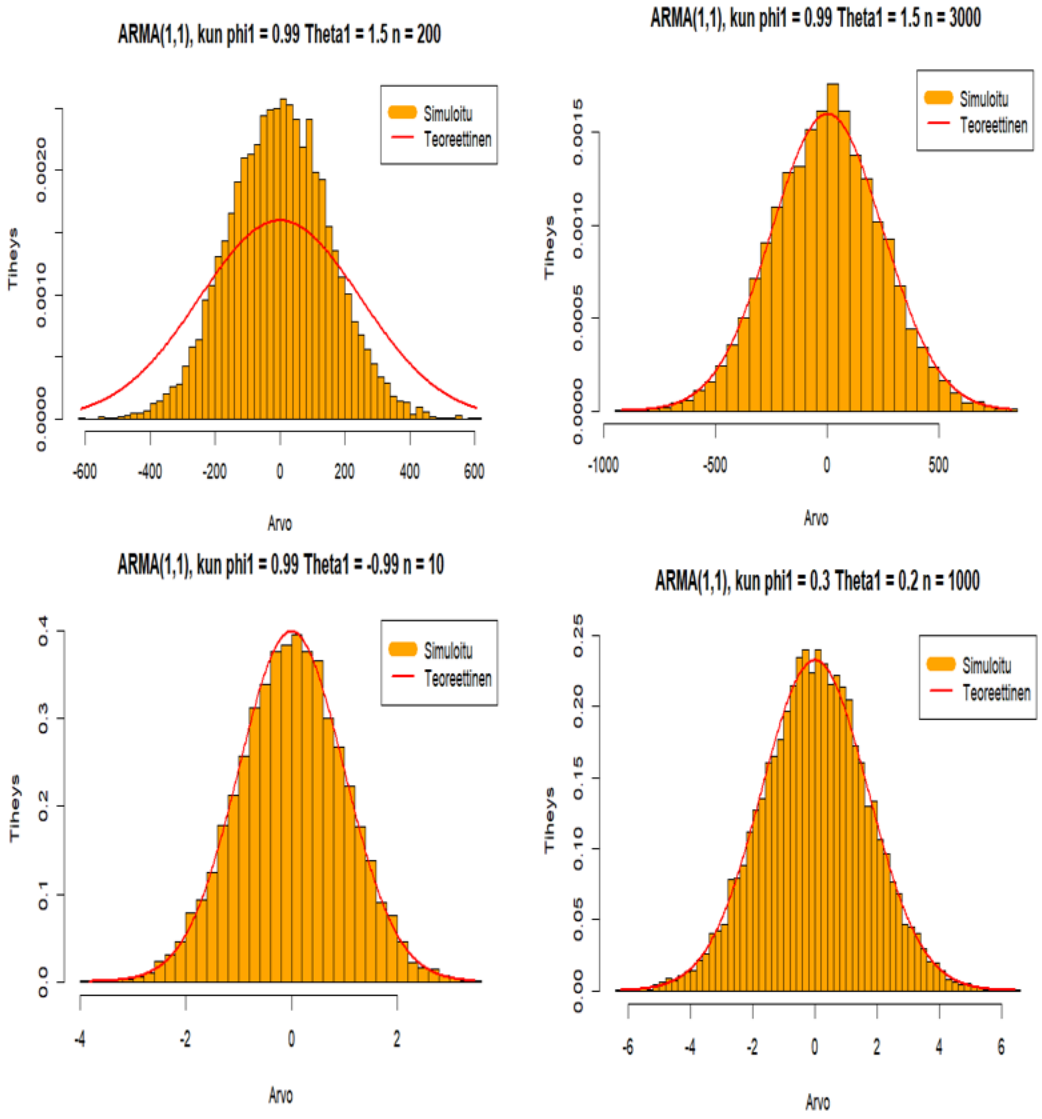
□

## 10.2 ARMA(1,1)-KRL normaaliapproksimaatiossa

ARMA-mallissa MA-osa perustuu käännettävyysehtoon, joka ARMA(1,1)-mallin tapauksessa perustuu aikaisemmin käsiteltyyn MA(1)-prosessin käännettävyysehtoon (14). MA-prosessi on aina stationaarinen. Suuret parametrin arvot kuitenkin heikentävät keskiarvon jakauman lähestymistä kohti normaalijakaumaa, joten sillä on suuri vaikutus, kun tarkastellaan parametrien käyttäytymistä KRL:n kannalta.

Kuva 19 seuraavalla sivulla havainnollistaa tilannetta valituilla parametrien arvoilla. Ylhäällä vasemmalla esitetään tapaus, jossa MA(1)-prosessin kääntyvyysehto ei täyty, sillä $\theta_1 > 1$. Samalla AR(1)-prosessin parametri on hyvin lähellä stationaarisuuden raja-arvoa, $\phi_1 = 0.99 < 1$. Kun otoskoko on $n = 200$, teoreettinen normaalijakauma ei vielä kuvaa havaittua jakaumaa tarkasti, eikä normaaliapproksimaatio ole hyvä. Lisäksi otoskeskiarvon asymptoottinen varianssi on tällöin hyvin suuri. Sen sijaan otoskoon kasvaessa ja saavuttaessa arvon $n = 3000$, asymptoottiset tulokset alkavat päteä, ja teoreettinen normaalijakauma vastaa jo hyvin havaittua normaaliapproksimaatiota.

Alhaalla vasemmalla puolestaan on tilanne, jossa AR(1)-prosessin parametri saa arvon $\phi_1 = 0.99 < 1$, jolloin malli on stationaarinen mutta hyvin lähellä stationaarisuuden raja-arvoa. MA(1)-prosessin parametri saa puolestaan arvon $-0.99$, mikä vastaa kääntyvyyden rajatapausta. Näillä parametrien arvoilla otoskeskiarvon asymptoottinen varianssi saa arvon 1, ja normaaliapproksimaatio on jo hyvin tarkka pienelläkin otoskoolla, $n = 10$. Mikäli parametrien arvot valitaan molempien mallien kannalta selvästi suotuisiksi, teoreettinen normaalijakauma seuraa erittäin tiiviisti simulaatiosta saatua keskiarvojen jakaumaa jo hyvin pienillä otoksilla. Havainnollistamisen vuoksi tähän on valittu otoskoko $n = 1000$, jolloin simulaatiosta saadut ostoskeskiarvojen jakaumat piirtyvät erittäin tiiviisti teoreettisen normaalijakauman sisälle.




Kuva 19: ARMA(1,1)-KRL eri parametrien $\phi_1$ ja $\theta_1$ arvoilla

Mikäli AR-osan stationaarisuus ei toteutuisi, voitaisiin käyttää ARIMA$(p, d, q)$-mallia, jossa differensiointiaste $d$ tekee prosessista stationaarisen. Tämä kuitenkin edellyttäisi huolellista mallinvalintaa ja useiden eri tekijöiden huomioon ottamista. Koska ARIMA$(p, d, q)$-malli ei ole tämän työn kannalta oleellinen, niin aihe jätetään käsittelemättä. Tämä on kuitenkin selitetty kompaktisti lähteessä (*[34] s. 134-144*).

# 11 Simulaatiosovelluksia

Aikaisemmissa luvuissa simulaatiot keskittyvät asymptoottisen jakauman tarkasteluun normaaliapproksimaatioiden kautta vaihtelevilla parametrien suuruuksilla. Näissä käytettiin vakiona odotusarvon $\mu = 0$ ja varianssin $\sigma^2 = 1$ arvoja. Tässä osiossa tarkastellaan aikaisemmin esitettyjen aikasarjamallien 95 %:n luottamusvälin peittotodennäköisyyksiä eri tilanteissa, kuten vaihtelevilla varianssin ja otoskoon arvoilla. Tavoitteena on havainnollistaa, miten mallit käyttäytyvät asymptoottisesti, kun $n \to \infty$. Tulokset esitetään kuvien avulla, ja malleja verrataan keskenään selkeyden ja havainnollisuuden lisäämiseksi. Jokainen simulaatio on toteutettu 10 000 simulaation ajolla.

Peittotodennäköisyydet tarkoittavat sitä osuutta simulaatioista, joissa rakennettu luottamusväli sisältää parametrin todellisen arvon. Jos luottamusväli on nimellisesti 95 %:n luottamusväli, ideaalitilanteessa peittotodennäköisyys on 0.95. Luottamusvälit perustuvat estimointisuureen asymptoottiseen normaalijakaumaan. Asymptoottinen varianssi $Var_{Asymp}$ määritellään mallikohtaisesti ja se kuvaa estimaattorin varianssia, kun otoskoko $n \to \infty$. Asymptoottinen keskihajonta on tämän neliöjuuri

$$Sd_{Asymp} = \sqrt{Var_{Asymp}}.$$

95 %:n luottamusväli on puolestaan muotoa

$$\bar{X} \pm z_{0.975} \frac{Sd_{Asymp}}{\sqrt{n}}, \tag{28}$$

jossa $z_{0.975}$ on standardinormaalijakauman 97.5 %:n kvantiili. Simulaatioissa peittotodennäköisyys estimoidaan laskemalla, kuinka usein tämä luottamusväli sisältää tunnetun odotusarvon $\mu = 0$. Keskihajonta kullekin osiossa käytetylle mallille on laskettu seuraavasti

$$\text{AR(2)-KRL } \; Var_{Asymp} = \frac{\sigma^2}{(1-\phi_1-\phi_2)^2}, \; Sd_{Asymp} = \frac{\sigma}{|1-\phi_1-\phi_2|},$$

$$\text{MA(2)-KRL } \; Var_{Asymp} = \sigma^2(1+\theta_1+\theta_2)^2, \; Sd_{Asymp} = \sigma|1+\theta_1+\theta_2|,$$

$$\text{ARMA(1,1)-KRL } \; Var_{Asymp} = \sigma^2 \left(\frac{1+\theta_1}{1-\phi_1}\right)^2, \; Sd_{Asymp} = \sigma \left|\frac{1+\theta_1}{1-\phi_1}\right|.$$

Näissä 95 %:n luottamusvälin laskemiseen on käytetty yllä olevaa kavaa (28).

## 11.1 Peittotodennäköisyyksiä eri $\sigma^2$:n ja $n$:n arvoilla

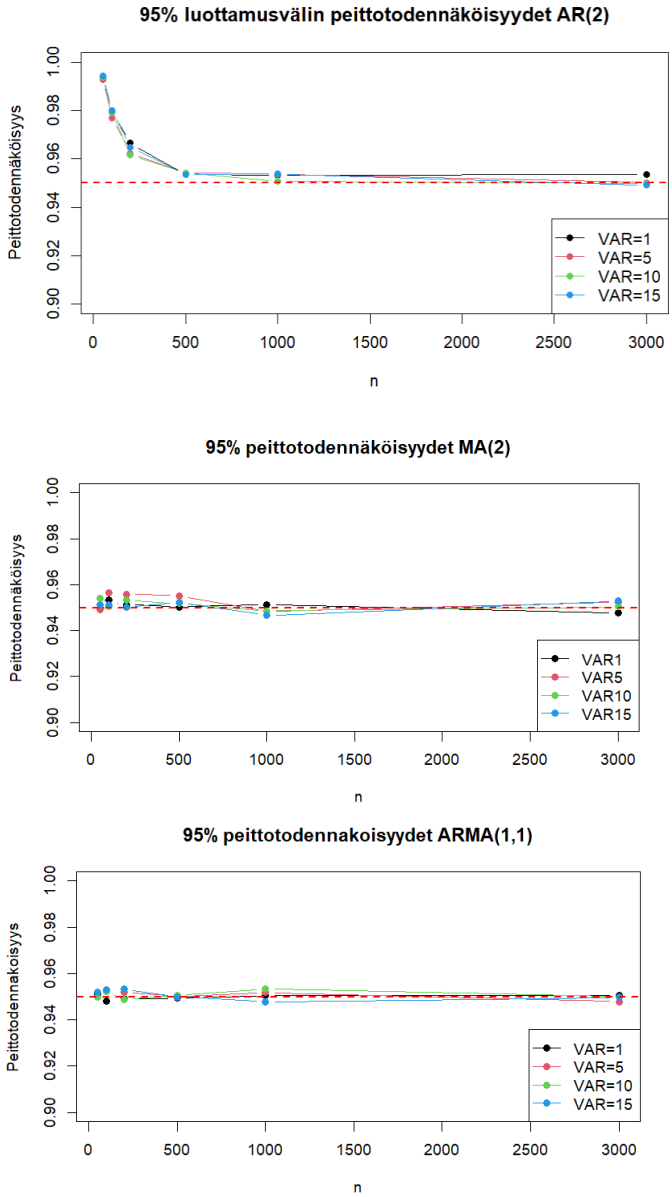
Seuraavassa on esitetty tilanne, jossa on tehty 10 000 simulaatiota malleille AR(2)-MA(2) ja ARMA(1,1) eri varianssin ja otoskoon arvoilla seuraavasti

$$\begin{aligned} \text{Otoskoko } n &= 50, 100, 200, 500, 1000, 3000 \\ \text{Varianssi } \sigma^2 &= 1, \; 5, \; 10, \; 15 \\ \text{Odotusarvo } \mu &= 0 \end{aligned}$$

Kaikkien mallien parametrien arvot ovat samat ja ne on valittu seuraavasti

$$\begin{aligned} \text{AR(2)-KRL} \quad & \phi_1 = 0.2, \ \phi_2 = 0.7 \\ \text{MA(2)-KRL} \quad & \theta_1 = 0.2, \ \theta_2 = 0.7 \\ \text{ARMA(1,1)-KRL} \quad & \phi_1 = 0.2, \ \theta_1 = 0.7 \end{aligned}$$

Tavoitteena kokeessa on tutkia vaikuttaako varianssin koko peittotodennäköisyyksiin ja millä otoskoon suuruudella tavoiteltu 95 %:n luottamusväli saavutetaan. Kuvissa punainen poikkiviiva kuvaa 95 %:n luottamusväliä.




Kuva 20: AR(2)-KRL, MA(2)-KRL ja ARMA(1,1)-KRL:n peittotodennäköisyydet.

Kuvassa 20 huomaa, että AR(2)-KRL:n 95 %:n luottamusvälit peittotodennäköisyyksille ovat ylioptimistisia pienillä otoksilla, mutta paranevat, kun $n \to \infty$. Kun otoskoko

lähestyy arvoa $n = 500$, peittotodennäköisyydet alkavat saavuttamaan 95 %:n tavoitetasoa. Kuvissa parametrien arvot ovat lähellä stationaarisuuden raja-arvoa 1, niiden saadessa arvon $\phi_1 + \phi_2 = 0.90$, joka osaltaan vaikuttaa peittotodennäköisyyksiin. Myöhemmin kuvassa 21 huomataan, että tilanne muuttuu paremmaksi, kun parametrien $\phi_1$ ja $\phi_2$ arvot ovat pienempiä. Varianssin suuruudella ei näyttäisi tässä olevan vaikutusta.

ARMA(1,1)-prosessi puolestaan pystyy mallintamaan sekä lyhytaikaisia, että pitkäaikaisia riippuvuuksia, jolloin vastaavasti KRL:n peittotodennäköisyydet ovat hyviä jo alusta alkaen. MA(2)-prosessi puolestaan riippuu edellisten virheiden vaikutuksesta, jolloin se osuu useammin oikeaan ja peittotodennäköisyys on parempi. Tässäkin peittotodennäköisyydet pysyvät lähes samoina riippumatta varianssin arvosta. Varianssin arvoilla ei siten näyttäisi olevan vaikutusta mallien peittotodennäköisyyksiin. Seuraavassa osiossa tilannetta sovelletaan vielä eri varianssin arvoilla tapauksessa, joissa AR(2)-prosessin varianssi on määritelty Yule-Walker-menetelmään (8) perustuen.

## 11.2 AR(2) peittotodennäköisyydet Yule-Walker-varianssilla

Kuten luvussa 8 todettiin, AR(2)-prosessin varianssi perustuu autokovarianssien ja autokorrelaatiotioiden viiveoperaattoreihin, jotka estimoidaan käytännössä niin sanotulla Yule-Walker-menetelmällä. Seuraavaksi tarkastellaan AR(2)-tilannetta, jossa asymptoottisen varianssin parametrit $\phi_1,\ \phi_2,\ \sigma^2$ eivät ole ennalta tunnettuja, vaan ne estimoidaan aineistosta Yule-Walker-yhtälöistä (8) johdetulla varianssilla ja siitä saadulla estimointipohjaisella asymptoottisella keskihajonnalla, joka arvioidaan käyttämällä Yule-Walker estimaatteja $\hat{\phi_1},\ \hat{\phi_2},\ \hat{\sigma^2}$ seuraavasti

$$\hat{Sd_{Asymp}} = \frac{\hat{\sigma}}{|1 - \hat{\phi_1} - \hat{\phi_2}|}.$$

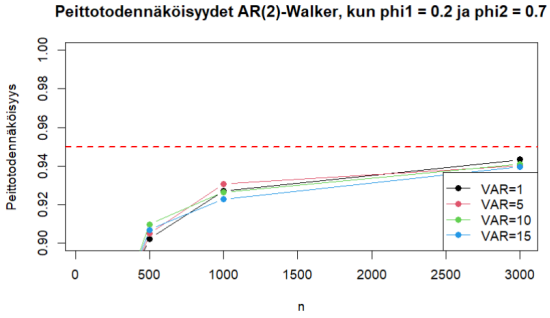
95 %:n luottamusväli puolestaan saadaan, kuten aikaisemmin, mutta nyt käytetään estimointipohjaista keskihajontaa

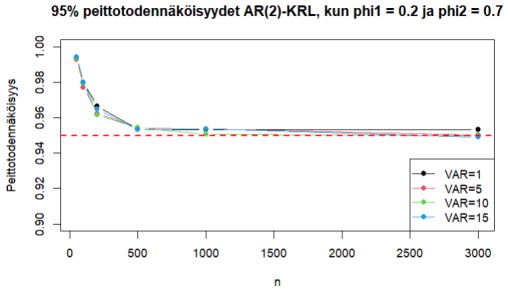
$$\bar{X} \pm z_{0.975} \frac{\hat{Sd_{Asymp}}}{\sqrt{n}}.$$

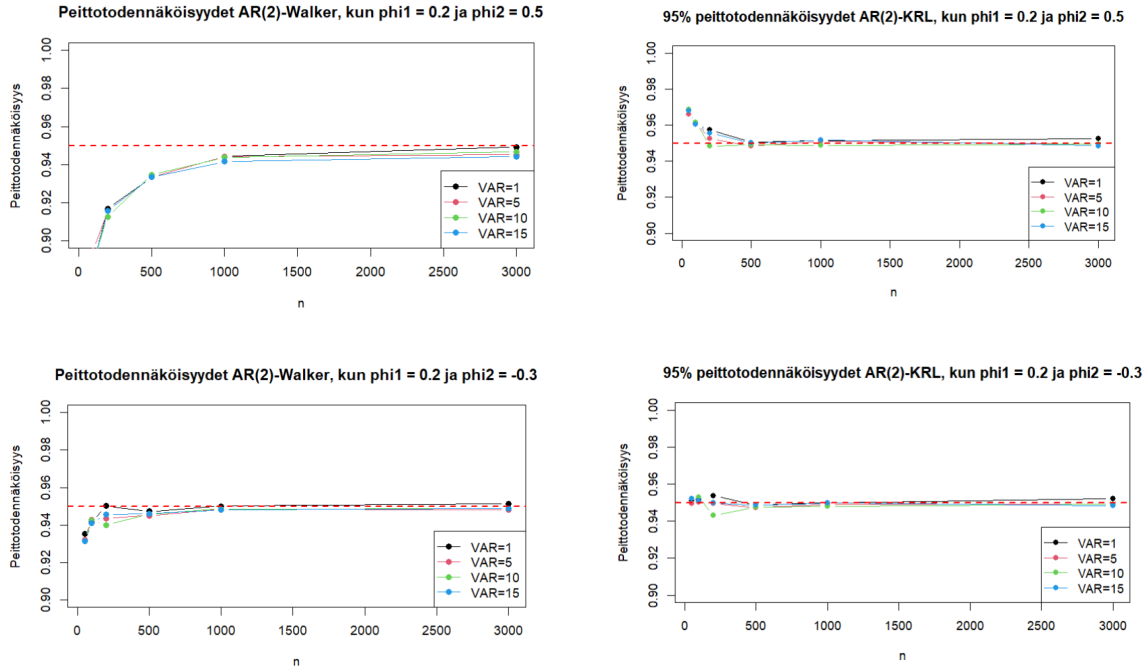
Varianssin ja otoskoon arvot ovat samat kuin edellisessä luvussa. Mallien parametrien arvot puolestaan vaihtelevat ja ovat seuraavat

$$(\phi_1 = 0.2,\ \phi_2 = 0.7),\ (\phi_1 = 0.2,\ \phi_2 = 0.5),\ (\phi_1 = 0.2,\ \phi_2 = -0.3)$$

Vertailun vuoksi rinnalle on otettu mukaan myös tapaukset, joissa parametrien arvot oletetaan tunnetuiksi. Nämä ovat kuvissa oikealla puolella.






Kuva 21: AR(2), Yule-Walker-peittotodennäköisyydet vs. Ennalta tunnetut parametrien arvot.

Kuvan 21 ylempi paneeli edustaa tilannetta, jossa parametrien arvot ovat samat kuin kuvassa 20, jossa liikutaan lähellä stationaarisuuden raja-arvoa 1. Parametrit saavat tällöin arvon $\phi_1 + \phi_2 = 0.90$, jolloin luottamusvälien peittotodennäköisyydet jäävät alle tavoitellun 95 %:n välin, eivätkä ne kata todellista keskiarvoa niin usein kuin pitäisi. Vasta otoskoon ollessa $n = 3000$ alkavat peittotodennäköisyydet saavuttamaan 95 %:n luottamustasoa. Tässäkin voidaan luottaa asymptoottisiin tuloksiin, kun $n \to \infty$, jolloin peittotodennäköisyydet alkavat olemaan halutulla tasolla. Kuvan 21 keskipaneeli puolestaan edustaa tilannetta, jossa $\phi_1 + \phi_2 = 0.70$, jolloin luottamusvälit alkavat saamaan parempia peittotodennäköisyyksiä jo pienemmillä otoksilla. Alin paneeli puolestaan edustaa tilannetta, jossa $\phi_1 = 0.2$ ja $\phi_2 = -0.3$, jolloin AR(2)-prosessin polynomien karakteristiset juuret (19) saadaan laskettua toisen asteen polynomin ratkaisukaavalla seuraavasti

$$
\begin{aligned}
&1 - \phi_1 z - \phi_2 z^2 = 0 \implies 1 - 0.2z - (-0.3)z^2 = 0 \implies 0.3z^2 - 0.2z + 1 = 0 \\
&a = \phi_2 = 0.3,\ b = \phi_1 = -0.2,\ c = 1 \\
&\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{0.2 \pm \sqrt{-1.16}}{0.6} = \frac{0.2 \pm i\sqrt{1.16}}{0.6} \\
&|z| = \sqrt{(0.2/0.6)^2 + (\sqrt{1.16}/0.6)^2} = 1.83 > 1
\end{aligned}
$$

Koska $|z| > 1$, niin polynomien karakteristiset juuret sijaistevat yksikköympyrän ulkopuolella. Koska myös ehdot (20) pätevät, niin malli on stationaarinen. Tällöin peittotodennäköisyydet saavuttavat 95 %:n luottamusvälin jo hyvin pienillä otoksilla. Todellisessa tilanteessa parametreja ei tiedetä etukäteen, vaan ne pitää estimoida aineistosta.

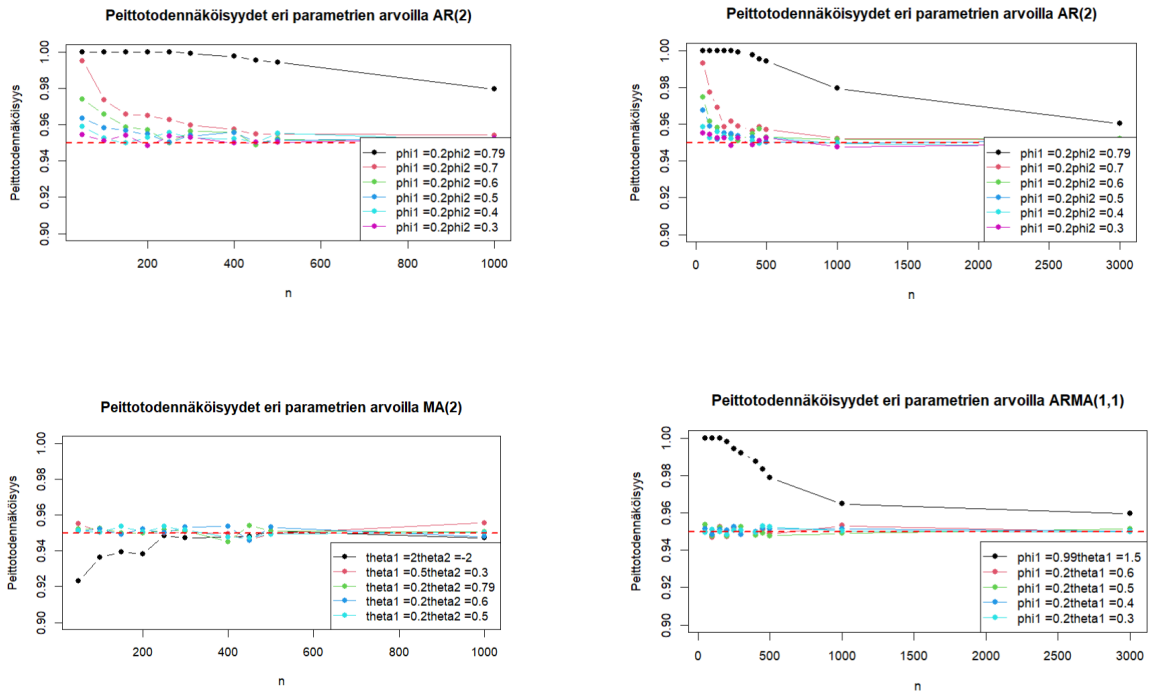
Tämä tekee luottamusvälien muodostumisesta haastavampaa, etenkin pienillä otoksilla. Kun parametrit on estimoitu AR(2)-mallin luottamusvälit eivät välttämättä kata todellista keskiarvoa halutulla tasolla, varsinkaan lähellä stationaarisuuden rajaa ja pienissä aineistoissa. Suurilla otoksilla asymptoottiset tulokset yleensä pätevät hyvin. Siten AR(2)-malli voi toimia luotettavasti, mutta pientä aineistoa käyttäessä on syytä olla varovainen ja tiedostaa, että parametrien estimointi vaikuttaa epävarmuuteen.

Vertailun vuoksi kuvassa 21 esitetään tulokset myös tilanteissa, joissa prosessin parametrit $\phi_1,\ \phi_2,\ \sigma^2$ eivät ole estimoitu suoraan aineistosta, vaan ne on oletettu tunnetuiksi ja käytetty sellaisenaan luottamusvälien laskennassa. Näissä tapauksissa luottamusvälit on muodostettu suoraan aikaisemmin osiossa 8.1 johdetun AR(2)-mallin asymptoottisen varianssin perusteella. Tässä tilanteessa luottamusvälien peittotodennäköisyydet ovat lähellä tavoiteltua 95 %:n tasoa jo pienillä otoksilla. Tämä osoittaa, että kun parametrien arvot tunnetaan, AR(2)-mallin luottamusvälit toimivat tarkasti. Parametrien $\phi_1$ ja $\phi_2$ arvojen suuruus näyttää kuitenkin vaikuttavan tuloksiin. Kun niiden yhteisvaikutus $\phi_1 + \phi_2$ loittonee stationaarisuuden raja-arvon 1 läheisyydestä, myös Yule-Walker-menetelmällä estimoidut peittotodennäköisyydet alkavat saavuttamaan vastaavia tuloksia. Mielenkiintoista on tarkastella myös käyrien lähestymissuuntaa. Kun käytetään asymptoottista varianssia, peittotodennäköisyydet lähestyvät tavoitetasoa yläpuolelta ja ovat pienillä otoksilla usein ylioptimistisia. Tällöin luottamusväli voi olla niin leveä, että sen peittotodennäköisyys nousee lähelle trivaalin 100 %:n luottamusvälin tasoa, mikä ei enää ole informatiivista. Tämä johtuu siitä, että Asymptoottinen menetelmä perustuu oletukseen $n \to \infty$, mikä alkaa toteutua vasta suuremmilla otoksilla. Sen sijaan, kun parametrit estimoidaan suoraan Yule-Walker-menetelmällä, peittotodennäköisyydet lähestyvät tavoitetasoa alapuolelta ja ne antavat pienemmillä otoksilla konservatiivisempia tuloksia, koska autokovarianssit ja autokorrelaatiot lasketaan suoraan aineistosta Yule-Walker-menetelmällä (21).

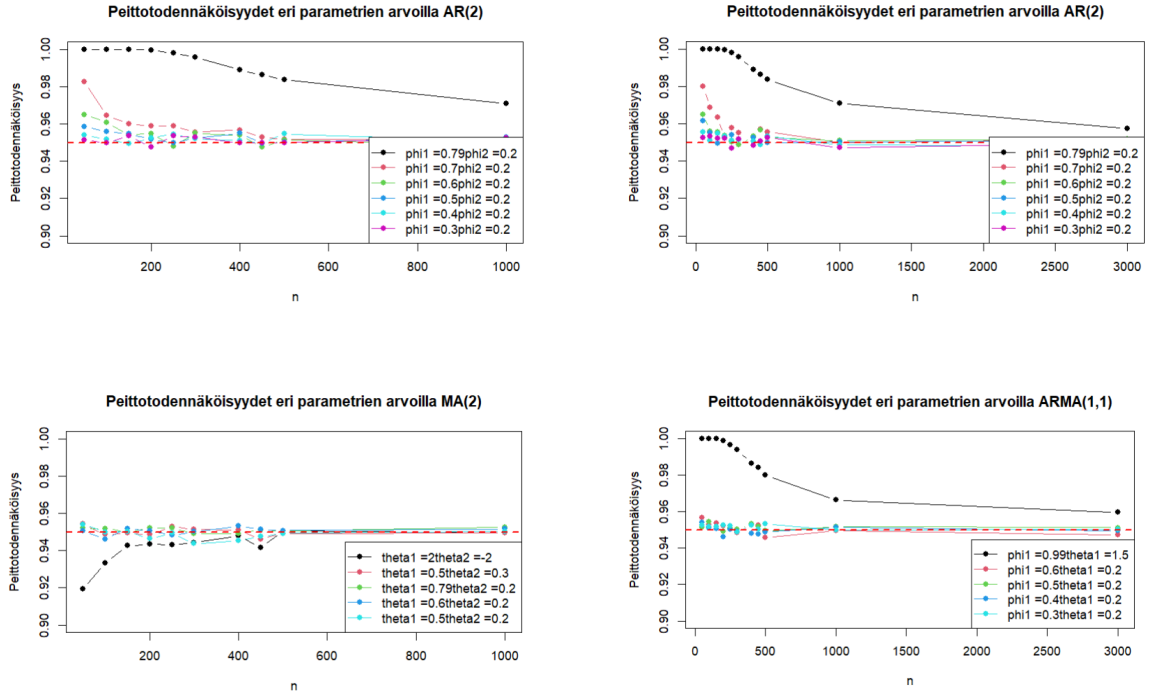
## 11.3 Peittotodennäköisyydet eri parametrien ja $n$:n arvoilla

Edellisessä osiossa kävi ilmi, että varianssin arvo ei vaikuta juurikaan peittotodennäköisyyksiin, joten tässä osiossa käytetään vakiovarianssia. Sen sijaan malleille AR(2), MA(2) ja ARMA(1,1) testataan vastaavaa tilannetta eri parametrien arvoilla. Kuvassa 22 ensimmäisen kertoimen arvo on 0.2 ja toisen kertoimen arvo kasvaa asteittain kohti ykköstä sen saadessa arvot 0.3-0.7. Kuva 23 esittää tilanteen päinvastoin, jossa ensimmäisen kertoimen arvo kasvaa asteittain välillä 0.3-0.7 ja toinen kerroin on vakio 0.2. Jokaiselle mallille kuvataan siis viisi eri tilannetta. Mallit on estimoitu seuraavilla arvoilla

$$\begin{aligned} \text{Otoskoko } n &= 50, 100, 150, 200, 250, 300, 400, 450, 500, 1000 \\ \text{Varianssi } \sigma^2 &= 5 \\ \text{Odotusarvo } \mu &= 0 \end{aligned}$$




Kuva 22: AR(2)-KRL, MA(2)-KRL ja ARMA(1,1)-KRL:n peittotodennäköisyydet eri parametrien arvoilla, kun ensimmäinen kerroin saa arvon 0.2




Kuva 23: AR(2)-KRL, MA(2)-KRL ja ARMA(1,1)-KRL:n peittotodennäköisyydet eri parametrien arvoilla, kun toinen kerroin saa arvon 0.2

Molemmissa kuvissa 22 ja 23 havaitaan, että AR(2)-KRL:ssä parametrien yhteenlaskettujen arvojen $(\phi_1 + \phi_2)$ lähestyessä stationaarisuuden raja-arvoa 1, ne saavuttavat

95 %:n peittotodennäköisyyden vasta silloin, kun otoskoko on noin 500. Lisäksi on tilanne, jossa liikutaan erittäin lähellä stationaarisuuden raja-arvoa 1, parametrien $\phi_1$ ja $\phi_2$ saadessa yhteisarvon $0.99 < 1$. Tällöin ne saavat ylioptimistisia peittotodennäköisyyksiä pienillä otoksilla, kunnes otoksien kasvaessa alkavat saavuttamaan haluttua 95 %:n peittotodennäköisyyttä. Kuitenkin on havaittavissa, että mitä pienempi parametrien yhteenlaskettu arvo on kyseessä, sitä aikaisemmin haluttu taso saavutetaan. Esimerkiksi kun parametrien arvot ovat $(\phi_1 + \phi_2 = 0.2 + 0.3 = 0.5 < 1)$, niin ne saavuttavat 95 %:n peittotodennäköisyyden jo otoskoon ollessa $n = 100$.

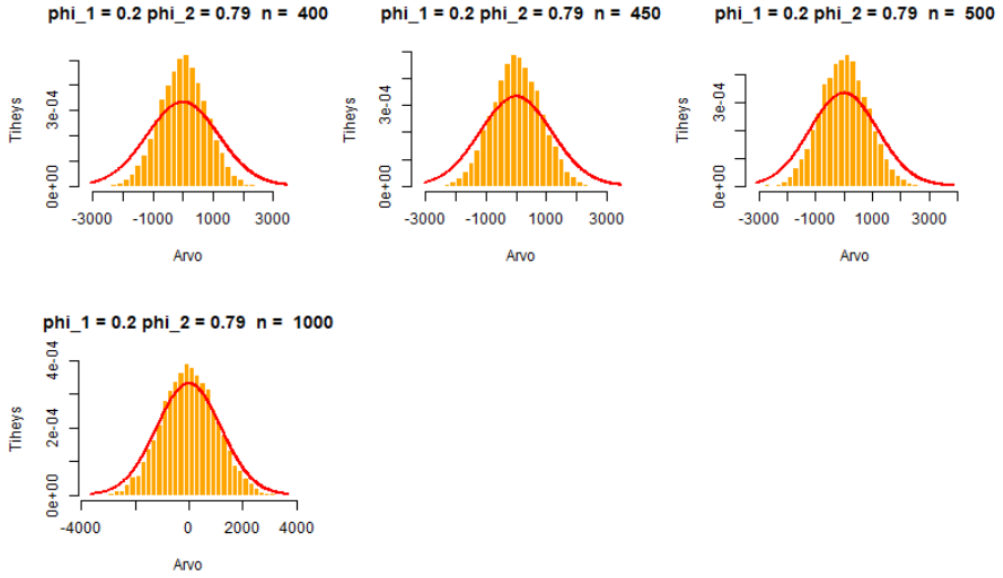
Kuvasta 23 on silmämääräisesti havaittavissa, että malli jossa $\phi_1$:n arvot kasvavat asteittain välillä 0.3-0.7 ja $\phi_2$ saa vakioarvon 0.2, peittotodennäköisyydet ovat hieman parempia pienemmillä otoksilla, kuin kuvassa 22, jossa $\phi_1$ saa vakioarvon 0.2 ja $\phi_2$:n arvot kasvavat asteittain välillä 0.3-0.7. Kaiken kaikkiaan kummastakin kuvasta voi päätellä, että suuret $\phi$:n arvot ovat kriittisempiä tulosten suhteen ja tällöin myös otoskoon tulee olla suurempi.

Sen sijaan MA(2)- ja ARMA(1,1)-KRL:t antavat hyviä tuloksia jo alusta saakka ja ne kattavat 95 %:n luottamusvälin suhteellisen hyvin myös pienillä otoksilla. Tarkemmin tarkasteltaessa myös näissä pienempi yhteenlaskettu arvo tuottaa hieman parempia tuloksia, mutta ero on lähes mitätön.

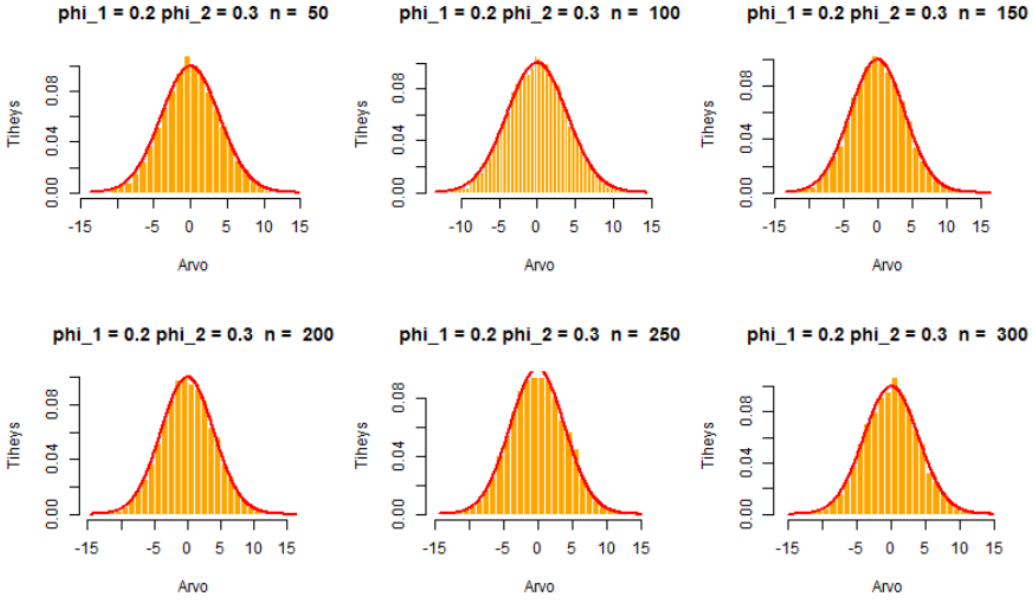
Kuvassa 22 ja 23 MA(2)-KRL:ään on tuotu mukaan osiossa 10.2 oleva tilanne, jossa kääntyvyysehto ei toteudu, parametrien saadessa arvot $\theta_1 = 2$ ja $\theta_2 = -2$. Vaikka MA-prosessi on aina stationaarinen, niin kääntyvyysehdon rikkoutumisella näyttäisi olevan vaikutusta nopeuteen. Asymptoottiset tulokset näyttäisivät kuitenkin toteutuvan hyvin kun $n$ kasvaa, jolloin se saavuttaa 95 %:n peittotodennäköisyyden kun otoskoko on noin $n = 500$. Osiosta 10.2 on tuotu mukaan myös tilanne, jossa parametrit ovat suotuisia ja kääntyvyys toteutuu hyvin parametrien saadessa arvot $\phi_1 = 0.5$ ja $\phi_2 = 0.3$. Kuvaajassa tätä esittää punainen jana ja sitä on vaikea havaita, koska se noudataa erittäin tarkasti punaisella katkoviivalla merkittyä 95 %:n luottamusväliä jo heti alusta saakka.

ARMA(1,1)-KRL:ään on tuotu mukaan osiossa 10.2 oleva tilanne, jossa $\phi_1 = 0.99$ ja $\theta_1 = 1.5$, jolloin AR(1)-osa on erittäin lähellä stationaarisuuden raja-arvoa 1 ja MA(2)-osa puolestaan ei ole kääntyvä. Tässä voidaan päätyä samaan lopputulemaan kuin osiossa 10.2, kun otoskoko on $n = 200$, niin peittotodennäköisyydet saavat ylioptimistisia arvoja. Tämä näkyi osiossa 10.2 sillä, että teoreettinen normaalijakauma ja normaaliapproksimaatio eivät kohdanneet ja normaalijakauma oli hyvin leveä. Tulokset alkoivat kuitenkin parantua, kun otoskoko kasvoi ja $n = 3000$ kohdalla simulaatiosta saadut keskiarvojen jakaumat alkoivat noudattamaan teoreettista normaalijakaumaa.

Kuvassa 24 seuraavalla sivulla esiintyvät normaaliapproksimaatiot esittävät kuvassa 22 olevaa AR(2)-tapausta, parametrien arvoilla $(\phi_1 = 0.2, \phi_2 = 0.79)$, jolloin liikutaan erittäin lähellä stationaarisuuden raja-arvoa 1. Tällöin normaaliapproksimaatio alkaa saavuttamaan teoreettista normaalijakaumaa vasta otoskoon ollessa noin $n = 1000$. Kuva 25 puolestaan esittää tilannetta, jossa $(\phi_1 = 0.2, \phi_2 = 0.3)$, jolloin normaaliapproksimaatio alkaa saavuttamaan teoreettista normaalijakaumaa jo pienemmillä otoksilla ja täydellisesti, kun $n = 150$.




Kuva 24: AR(2)-KRL eri parametrien arvoilla, kun $\phi_1 = 0.2$ ja $\phi_2 = 0.79$




Kuva 25: AR(2)-KRL eri parametrien arvoilla, kun $\phi_1 = 0.2$ ja $\phi_2 = 0.3$

## 11.4 Taulukot ääripäiden peittotodennäköisyyksistä

Seuraavassa esitetään taulukot peittotodennäköisyyksistä koskien osiota 11.3. Mukaan on otettu kuvan 22 tapaukset koska kuvan 23 tulokset eivät oleellisesti eronneet siitä. Taulukoihin on koottu vertailun vuoksi kunkin mallin ääripäiden arvoja, jotta erot näiden välillä tulisi selvimmin esille.

Taulukko 1: AR(2)-KRL peittotodennäköisyydet, kun $\phi_1 = 0.2$ ja $\phi_2 = 0.3$, $\phi_2 = 0.79$

| $\phi_1$ | $\phi_2$ | $n$ | Peittotodennäköisyys |
|---|---|---|---|
| 0.2 | 0.30 | 50 | 0.9551 |
| 0.2 | 0.30 | 100 | 0.9545 |
| 0.2 | 0.30 | 150 | 0.9527 |
| 0.2 | 0.30 | 200 | 0.9527 |
| 0.2 | 0.30 | 250 | 0.9482 |
| 0.2 | 0.30 | 300 | 0.9524 |
| 0.2 | 0.30 | 400 | 0.9488 |
| 0.2 | 0.30 | 450 | 0.9512 |
| 0.2 | 0.30 | 500 | 0.9527 |
| 0.2 | 0.30 | 1000 | 0.9477 |
| 0.2 | 0.30 | 3000 | 0.9497 |
| 0.2 | 0.79 | 50 | 1.0000 |
| 0.2 | 0.79 | 100 | 1.0000 |
| 0.2 | 0.79 | 150 | 1.0000 |
| 0.2 | 0.79 | 200 | 1.0000 |
| 0.2 | 0.79 | 250 | 1.0000 |
| 0.2 | 0.79 | 300 | 0.9992 |
| 0.2 | 0.79 | 400 | 0.9975 |
| 0.2 | 0.79 | 450 | 0.9953 |
| 0.2 | 0.79 | 500 | 0.9942 |
| 0.2 | 0.79 | 1000 | 0.9794 |
| 0.2 | 0.79 | 3000 | 0.9604 |

Taulukko 1 kuvaa AR(2)-KRL tilannetta, jossa pienet parametrien arvot $\phi_1 = 0.2, \phi_2 = 0.3$ saavat hyviä peittotodennäköisyyksiä jo alusta saakka. Kun taas erittäin lähellä stationaarisuuden raja-arvon 1 läheisyydessä liikkuvat parametrit $\phi_1 = 0.2, \phi_2 = 0.79$ saavat aluksi ylioptimistisia peittotodennäköisyyksiä, kunnes otoskoon suurentuessa ne alkavat lähestyä tavoiteltua 95 %:n tasoa.

Taulukko 2: MA(2)-KRL peittotodennäköisyydet, kun $\theta_1 = 2$ ja $\theta_2 = -2$, $\theta_2 = 0.79$

| $\theta_1$ | $\theta_2$ | $n$ | Peittotodennäköisyys |
|---|---|---|---|
| 2.0 | -2.00 | 50 | 0.9219 |
| 2.0 | -2.00 | 100 | 0.9345 |
| 2.0 | -2.00 | 150 | 0.9318 |
| 2.0 | -2.00 | 200 | 0.9459 |
| 2.0 | -2.00 | 250 | 0.9425 |
| 2.0 | -2.00 | 300 | 0.9448 |
| 2.0 | -2.00 | 400 | 0.9467 |
| 2.0 | -2.00 | 450 | 0.9524 |
| 2.0 | -2.00 | 500 | 0.9475 |
| 2.0 | -2.00 | 1000 | 0.9456 |
| 0.2 | 0.79 | 50 | 0.9502 |
| 0.2 | 0.79 | 100 | 0.9517 |
| 0.2 | 0.79 | 150 | 0.9517 |
| 0.2 | 0.79 | 200 | 0.9481 |
| 0.2 | 0.79 | 250 | 0.9481 |
| 0.2 | 0.79 | 300 | 0.9517 |
| 0.2 | 0.79 | 400 | 0.9537 |
| 0.2 | 0.79 | 450 | 0.9459 |
| 0.2 | 0.79 | 500 | 0.9502 |
| 0.2 | 0.79 | 1000 | 0.9508 |

Taulukko 2 kuvaa MA(2)-KRL tilannetta, jossa kääntyvyysehto ei toteudu parametrien arvoilla $\theta_1 = 0.2, \theta_2 = -2$. Tämä osoittautui kuvassa 22 siten, että jakauma oikeni hitaasti pienillä otoksilla, kunnes saavutti halutun 95 %:n tason noin $n = 300$ kohdalla.

Toisessa tilanteessa parametrit saavat arvot $\theta_1 = 0.2, \theta_2 = 0.79$. Tässä kääntyvyysehto toteutuu ja jakauma saa 95 %:n peittotodennäköisyyksiä jo alusta saakka.

Taulukko 3: ARMA(1,1)-KRL peittotodennäköisyydet, kun $\phi_1 = 0.2, \theta_1 = 0.6$ ja $\phi_1 = 0.99$, $\theta_1 = 1.5$

| $\phi_1$ | $\theta_1$ | $n$ | Peittotodennäköisyys |
|---|---|---|---|
| 0.20 | 0.6 | 50 | 0.9541 |
| 0.20 | 0.6 | 100 | 0.9503 |
| 0.20 | 0.6 | 150 | 0.9511 |
| 0.20 | 0.6 | 200 | 0.9512 |
| 0.20 | 0.6 | 250 | 0.9508 |
| 0.20 | 0.6 | 300 | 0.9525 |
| 0.20 | 0.6 | 400 | 0.9480 |
| 0.20 | 0.6 | 450 | 0.9496 |
| 0.20 | 0.6 | 500 | 0.9482 |
| 0.20 | 0.6 | 1000 | 0.9532 |
| 0.20 | 0.6 | 3000 | 0.9493 |
| 0.99 | 1.5 | 50 | 1.0000 |
| 0.99 | 1.5 | 100 | 1.0000 |
| 0.99 | 1.5 | 150 | 0.9999 |
| 0.99 | 1.5 | 200 | 0.9989 |
| 0.99 | 1.5 | 250 | 0.9961 |
| 0.99 | 1.5 | 300 | 0.9929 |
| 0.99 | 1.5 | 400 | 0.9861 |
| 0.99 | 1.5 | 450 | 0.9846 |
| 0.99 | 1.5 | 500 | 0.9803 |
| 0.99 | 1.5 | 1000 | 0.9638 |
| 0.99 | 1.5 | 3000 | 0.9550 |

Taulukko 3 kuvaa ARMA(1,1)-KRL tilannetta, jossa parametrien arvot ovat $\phi_1 = 0.2, \theta_1 = 0.6$, jolloin stationaarisuus ja kääntyvyysehto toteutuvat, jolloin jakauma saa hyviä peittotodennäköisyyksiä jo alusta saakka. Parametrien arvoilla $\phi_1 = 0.99, \theta_1 = 1.5$ AR-osa on erittäin lähellä stationaarisuuden raja-arvoa 1. MA-osa puolestaan ei ole kääntyvä. Tämä osoittautui kuvassa 22 siten, että jakauma oikeni hitaasti pienillä otoksilla, kunnes alkoi saavuttamaan haluttua 95 %:n tason noin $n = 1000$ kohdalla.

# 12 Päätelmät

Keskeinen raja-arvolause on keskeinen menetelmä, kun tarkastellaan mallien ja estimaattorien asymptoottista käyttäytymistä. Tämä mahdollistaa sen, että esimerkiksi aikasarjamalleissa, kuten AR(p), MA(q) ja ARMA(p,q) asymptoottisten luottamusvälien peittotodennäköisyydet lähestyvät oikeaa luottamusvälin tasoa suurilla otoksilla, koska otoskeskiarvojen jakauma lähestyy normaalijakaumaa.

Lyhyenä yhteenvetona voidaan todeta, että AR(2):n vastaavat estimaattorit osoittivat epävakaampaa käyttäytymistä pienillä otoksilla, erityisesti kun mallin parametrit lähestyivät stationaarisuuden rajaa. Suuremmilla otoksilla ja erityisesti kun $n > 500$ peitotodennäköisyydet lähestyivät haluttua 95 %:n tasoa. Parametrin arvot $\phi_1+\phi_2$ vaikuttivat merkittävästi peittotodennäköisyyksiin, ja niiden lähestyessä arvoa 1 peittotodennäköisyys jäi alle tavoitellun tason pienillä otoksilla. AR(2)-mallin parametrit estimoitiin myös Yule-Walker-menetelmällä. Tässä peittotodennäköisyydet olivat hillitympiä ja alittivat 95 %:n tavoitteen erityisesti pienillä otoksilla. MA(2)- ja ARMA(1,1)-KRL antoivat hyviä peittotodennäköisyyksiä jo pienillä otoksilla. ARMA(1,1)-malli kykenee kuvaamaan sekä pitkäkestoista autokorrelaatiota AR(1), että lyhytaikaisia häiriöitä MA(1), mikä lisää mallin joustavuutta. MA(2)- prosessi on puolestaan aina stationaarinen, mikä tukee sen vakautta koko tarkastelujakson ajan.

Varianssin arvo ei juurikaan vaikuttanut peittotodennäköisyyksiin. Mallit käyttäytyivät pääsääntöisesti samalla tavalla huolimatta siitä, oliko varianssi pieni vai suuri. Mallin parametrien muutokset sen sijaan vaikuttivat merkittävästi peittotodennäköisyyksiin. Kun $\phi_1 + \phi_2$ lähestyivät stationaarisuuden raja-arvoa 1, peittotodennäköisyydet muuttuivat paremmiksi vasta suuremmilla otoksilla. Pienemmillä $\phi_1 + \phi_2$ arvoilla peittotodennäköisyydet puolestaan saavuttivat tavoitetason pienemmillä otoksilla. Voidaankin todeta, että AR(2)-malli on herkempi otoskoon ja parametrien vaihteluille, kun taas MA(2)- JA ARMA(1,1)-mallit tarjoavat vakaampia tuloksia jo pienemmillä otoksilla. Erityisesti AR(2)-mallissa on syytä noudattaa varovaisuutta, mikäli parametrien arvot lähestyvät stationaarisuuden rajaa, sillä pieni otoskoko voi johtaa epäluotettaviin tuloksiin. Suuremmat otoskoot kuitenkin parantavat kaikkien mallien peittotodennäköisyyksiä.

# Viitteet


[1] Al Nosedal. ARMA Models. University of Toronto, 2019. https://mcs.utm.utoronto.ca/~nosedal/sta457/arma-models.pdf.

[2] Charles Zaiontz. Autoregressive Processes Basic Concepts. Real Statistics Using Excel, 2025. https://real-statistics.com/time-series-analysis/autoregressive-processes/autoregressive-processes-basic-concepts/autoregressive-process-proofs/.

[3] Florian Kölbl. Aggregation of AR(2) Processes. Tilastotieteen laitos. Grazin teknillinen yliopisto. Diplomityö, 2006.

[4] John A. Dodson. Estimating the ARMA Model. 2022. https://www-users.cse.umn.edu/~dodso013/docs/ARMA.pdf.

[5] Juha Alho,Elja Arjas, Esa Läärä, Pekka Pere Tilastotieteen sanasto Suomen Tilastoseura ry Helsinki, 2023.

[6] Mathematische Zeitschrift. Verlag Von Julius Springer, Berlin, 1922.

[7] Jacob Bernoulli. *Ars Conjectandi*. Impensis Thurnisiorum, Fratrum, Basel, 1713. (Alkuperäisteos). Säilytys. Bodleian Libraries, Oxford University.

[8] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, Inc, 1995.

[9] Michael Bradley. *Modern Mathematics 1900 to 1950*. Infobase Publishing, 2019.

[10] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, Department of Statistics, Colorado State University, Fort Collins, USA.

[11] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, 2nd edition, 1991.

[12] A. De Moivre. *The Doctrine of Chances: Or, a Method of Calculating the Probabilities of Events in Play*. 3rd edition. London, 1756.

[13] Murat Kulahchi Douglas C. Montgomery, Cheryl L. Jennings. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Inc, 2008.

[14] Hans Fischer. *A History of the Central Limit Theorem*. Springer, New York, Dordrecht, Heidelberg, London, 2010.

[15] Central Limit Theorem for $\mu$ in AR(1) Model. Financial Time Series Analysis, week 5. Lecture slides.

[16] Massimo Guidolin. Autoregressive Moving Average (ARMA) Models and Their Practical Applications. University of Bocconi. Italy, 2018. Lecture slides. https://didattica.unibocconi.it/mypage/dwload.php?nomefile=Lecture_notes_2-3_final20180204220724.pdf.

[17] Anders Hald. *A History of Probability and Statistics and Their Applications Before 1750*. A John Wiley & Sons, Inc, 2003.

[18] Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer, 2007.

[19] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[20] C. C. Heyde and E. Seneta. *Statisticians of the Centuries*. Springer-Verlag, New York, Inc., 2001.

[21] Judy L. Klein. *A History of Time Series Analysis, 1662–1938*. Cambridge University Press, 1997.

[22] Pierre-Simon Laplace. *Théorie Analytique des Probabilites*. Paris, 1820.

[23] Pierre-Simon Laplace. *Théorie Analytique des Probabilités*. Translation by Richard J. Pulskamp, Book 1. Department of Mathematics, Xavier University, Cincinnati, Ohio, 2021.

[24] Pierre-Simon Laplace. *Théorie Analytique des Probabilités*. Translation by Richard J. Pulskamp, Book 2. Department of Mathematics, Xavier University, Cincinnati, Ohio, 2021.

[25] E. L. Lehmann. *Elements of Large-sample Theory*. Springer, 1999.

[26] C.B. Merzbach, U.C. Boyer. *A History of Mathematics.* A John Wiley & Sons, Inc, 3rd edition, 2021.

[27] David D. Nolte. Department of Statistics, Harvard University. `https://galileo-unbound.blog/2020/10/06/the-bountiful-bernoulli-of-basel/`.

[28] Al Nosedal. The Moving Average Models MA(1) and MA(2). University of Toronto. Lecture slides. `https://mcs.utm.utoronto.ca/~nosedal/sta457/ma1-and-ma2.pdf`.

[29] Henri Nyberg. Moniulotteinen aikasarja-analyysi. Lecture notes. University of Turku, Department of Mathematics and Statistics.

[30] E. S. Pearson. Studies in the History of Probability and Statistics. xiv: Some Incidents in the Early History of Biometry and Statistics, 1890–94. Biometrika, volume 52, 3-18, 1965.

[31] R. L. Plackett. Karl Pearson and the Chi-Squared Test. International Statistical Review. Volume 51, 1983.

[32] James G. MacKinnon Russell Davidson. *Estimation and Inference in Econometrics*. Oxford University Press, 1993.

[33] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* John Wiley & Sons, Inc., 2001.

[34] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications with R Examples*. Fourth edition. Springer, 2016.

[35] Bing Sung. Translations from James Bernoulli. Department of Statistics, Harvard University, 2005.

[36] Sutori. Historia de la estadística. `https://www.sutori.com/en/story/historia-de-la-estadistica--mtKBTrPQqmgZ1bFQKmgBPTuy`.

[37] E. D. Sylla. *Tercentenary of Ars Conjectandi. Jacob Bernoulli and the Founding of Mathematical Probability*. North Carolina State University, Raleigh, North Carolina, USA, 2006.

[38] Suomen tilastoseuran vuosikirja. Tommi Härkänen, 2024. `https://tilastoseura.fi/sites/tilastoseura.fi/files/2025-03/sts_vuosikirja_2024.pdf`.

[39] Leo Törnqvist. *Aikasarjojen analyysi ja ennustaminen*. Oy Gaudeamus Ab, 1974.

[40] Graham Upton and Ian Cook. *A Dictionary of Statistics*. Oxford University Press, 2014.

[41] William W.S. Wei. *Time Series Analysis. Univariate and Multivariate Methods*. Pearson Addison Wesley, 2006.

[42] Helsingin yliopisto. Ylioppilasmartikkeli 1853-1899.

[43] G. U. Yule. *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*. Philosophical Transactions of the Royal Society of London, 1927.

[44] G. U. Yule. *An Introduction to the Theory of Statistics*. C. Griffin, 1936.

# Liitteet

# A Työssä käytetyt R-koodit

## A.1 Koodit ennen simulaatiota

### A.1.1 AR(1)-KRL

```
# Funktio AR(1)-prosessin generointiin
generointi_ar1 <- function(n, phi, sigma = 1, mu = 0) {
  x <- numeric(n)
  x[1] <- rnorm(1, mean = mu, sd = sigma / sqrt(1 - phi^2)) # Ensimmainen arvo (stationaariselle)
  for (t in 2:n) {
    x[t] <- mu + phi * (x[t - 1]-mu) + rnorm(1, mean = 0, sd = sigma)
  }
  return(x)
}

set.seed(10)
# Simulaation parametrit
n <- 100
m <- 10000
phi <- 0.7
mu <- 0

# sqrt(n) (x_hat - mu) jokaiselle simulaatiolle
simulaatio <- numeric(m)
for (i in 1:m) {
  x <- generointi_ar1(n, phi, mu = mu)
  x_hat <- mean(x)
  simulaatio[i] <- sqrt(n) * (x_hat - mu)
}

# Histogrammi
hist(simulaatio, breaks = 50, probability = TRUE,
     main = paste("Simuloitu Vs. Teoreettinen, kun phi=", phi, "ja n=", n),
     xlab = "Arvo",
     ylab = "Tiheys",
     col = "orange")

# Teoreettinen normaalijakauma
x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
y_arvo <- dnorm(x_arvo, mean = 0, sd = sqrt(1/(1 - phi)^2)) # Standardoitu normaali
lines(x_arvo, y_arvo, col = "red", lwd = 2)


legend("topright", legend = c("Simuloitu", "Teoreettinen"),
       col = c("orange", "red"), lwd = c(10, 2))

set.seed(10)

# Yksi AR(1) sarja
x_sample <- generointi_ar1(n, phi, mu = mu)

# Aikasarja
plot(1:n, x_sample, type = "l", col = "orange",
     main = paste("AR(1)-prosessi ajassa, phi =", phi, "ja n=", n),
     xlab = "Aika", ylab = "X_t")
```

### A.1.2 AR(2)-KRL

```
# Funktio AR(2)-prosessin generointiin
generointi_ar2 <- function(n, phi1, phi2, sigma = 1, mu = 0) {
  x <- numeric(n)

  # Käytetään tarkaa varianssin arvoa, joka saatu laskemalla Yule-Walkerilla
```

```r
  gamma0 <- (sigma^2 * (1-phi2))/ ((1-phi1^2-phi2^2) * (1-phi2) -2 * phi1^2 * phi2)
  # Alkuarvot stationaarisuuden säilyttämiseksi
  x[1] <- rnorm(1, mean = mu, sd = sqrt(gamma0))
  x[2] <- rnorm(1, mean = mu, sd = sqrt(gamma0))

  for (t in 3:n) {
    x[t] <- mu + phi1 * (x[t - 1] - mu) + phi2 * (x[t - 2] - mu) + rnorm(1, mean = 0, sd = sigma)
  }
  return(x)
}

set.seed(10)
# Simulaation parametrit
n <- 1000
m <- 10000
phi1 <- 0.3
phi2 <- 0.3
mu <- 0

# sqrt(n) (x_hat - mu) jokaiselle simulaatiolle
simulaatio <- numeric(m)
for (i in 1:m) {
  x <- generointi_ar2(n, phi1, phi2, mu = mu)
  x_hat <- mean(x)
  simulaatio[i] <- sqrt(n) * (x_hat - mu)
}

# Histogrammi
hist(simulaatio, breaks = 50, probability = TRUE,
     main = paste("Simuloitu Vs. Teoreettinen AR(2), kun phi_1 =", phi1, "phi_2 =", phi2, "ja n =", n),
     xlab = "Arvo",
     ylab = "Tiheys",
     col = "orange")

# Teoreettinen normaalijakauma
x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
y_arvo <- dnorm(x_arvo, mean = 0, sd = sqrt(1 / (1 - phi1 - phi2)^2)) # Standardoitu normaali
lines(x_arvo, y_arvo, col = "red", lwd = 2)

legend("topright", legend = c("Simuloitu", "Teoreettinen"),
       col = c("orange", "red"), lwd = c(10, 2))

set.seed(10)

# Yksi AR(2) sarja
x_sample <- generointi_ar2(n, phi1, phi2, mu = mu)

# Aikasarja
plot(1:n, x_sample, type = "l", col = "orange",
     main = paste("AR(2)-prosessi ajassa, phi1 =", phi1, "phi2 =", phi2, "ja n=", n),
     xlab = "Aika", ylab = "X_t")
```

### A.1.3 MA(2)-KRL

```r
# Funktio MA(2)-prosessin generointiin
generointi_ma2 <- function(n, theta1, theta2, sigma = 1, mu = 0) {
  Z <- rnorm(n + 2, mean = 0, sd = sigma)
  X <- mu + Z[3:(n+2)] + theta1 * Z[2:(n+1)] + theta2 * Z[1:n]
  return(X)
}


set.seed(10)
# Simulaation parametrit
n <- 10
m <- 10000
theta1 <- 2
theta2 <- -2
mu <- 0
```

```r
# sqrt(n) (x_hat - mu) jokaiselle simulaatiolle
simulaatio <- numeric(m)
for (i in 1:m) {
  x <- generointi_ma2(n, theta1, theta2, mu = mu)
  x_hat <- mean(x)
  simulaatio[i] <- sqrt(n) * (x_hat - mu)
}


# Histogrammi
hist(simulaatio, breaks = 50, probability = TRUE,
     ylim = c(0, ymax),
     main = paste("Simuloitu Vs. Teoreettinen MA(2), kun theta_1 =", theta1, "theta_2 =", theta2, "ja n =", n),
     xlab = "Arvo",
     ylab = "Tiheys",
     col = "orange")

# Teoreettinen normaalijakauma
x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
v2 <- (1 + theta1 + theta2)^2 # asymptoottinen varianssi kun sigma^2=1
y_arvo <- dnorm(x_arvo, mean = 0, sd = sqrt(v2))

lines(x_arvo, y_arvo, col = "red", lwd = 2)

legend("topright", legend = c("Simuloitu", "Teoreettinen"),
       col = c("orange", "red"), lwd = c(10, 2))


set.seed(10)

# Yksi MA(2) sarja
x_sample <- generointi_ma2(n, theta1, theta2, mu = mu)

# Aikasarja
plot(1:n, x_sample, type = "l", col = "orange",
     main = paste("MA(2)-prosessi ajassa, theta_1 =", theta1, ", theta_2 =", theta2),
     xlab = "Aika", ylab = "X_t")
```

### A.1.4 De Moivre esimerkki

```r
    n <- 3600
p <- 0.5
mu <- p
sigma <- sqrt(p * (1 - p))
sqrt_n <- sqrt(n)
odotus_lkm <- n * p

# Poikkeama (1/2 * sqrt(n))
poikkeama <- 0.5 * sqrt_n

# Välin rajat onnistumismäärinä
ala_lkm <- odotus_lkm - poikkeama
yla_lkm <- odotus_lkm + poikkeama

# Suhteelliset osuudet
ala_osuus <- ala_lkm / n
yla_osuus <- yla_lkm / n

# Z-arvot normaalijakauman käyttamistä varten
z_ala <- (ala_osuus - mu) / (sigma / sqrt_n)
z_yla <- (yla_osuus - mu) / (sigma / sqrt_n)
z_ala
z_yla

# Simulointi keskiarvot
simulointi_maara <- 10000
set.seed(2025)
bar_X <- replicate(simulointi_maara, {
  mean(rbinom(n, size = 1, prob = p))
})
```

```
m <- 10000
simulaatio <- numeric(m)

# Simulointi KRL sqrt(n)(bar(X)-mu)/sigma
for (i in 1:m) {
  x <- rbinom(n, size = 1, prob = p)  # 3600 Bernoulli-kokeen tulokset
  x_hat <- mean(x)
  z_arvo <- sqrt(n) * (x_hat - mu) / sigma  # Z-arvo
  simulaatio[i] <- z_arvo
}

# Kuinka moni bar_X osuu väliin
osumia <- sum(bar_X >= ala_osuus & bar_X <= yla_osuus)
todennakoisyys_simulaatiosta <- osumia / simulointi_maara

cat("Analyyttinen todennäköisyys normaalijakaumasta:",
    round(pnorm(z_yla) - pnorm(z_ala), 6), "\n")
cat("Simuloitu todennäköisyys:", round(todennakoisyys_simulaatiosta, 6), "\n")

par(mfrow = c(1, 2))

# Ensimmäinen histogrammi (Otoskeskiarvot)
hist(bar_X, breaks = 50, col = "grey", probability = TRUE,
     main = "Otoskeskiarvot (n = 3600)", xlab = expression(bar(X)), ylab = "Tiheys",
     xlim = c(ala_osuus - 0.01, yla_osuus + 0.01))  # Rajataan x-akseli välin ympärille

# Normaalijakauma kuva 1
curve(dnorm(x, mean = mu, sd = sigma / sqrt_n),
      col = "red", lwd = 2, add = TRUE)

# Rajat kuva 1
abline(v = c(ala_osuus, yla_osuus), col = "black", lwd = 2, lty = 2)

legend("topright", legend = c("suhteelliset välit", "Normaalijakauma"),
       col = c("black", "red"), lty = c(2, 1), lwd = 2)

# Toinen histogrammi (KRL-simulaatio ja normaalijakauma)
hist(simulaatio, breaks = 50, probability = TRUE,
     main = paste("KRL-simulaatio-De Moivren mukaan, n =", n),
     xlab = expression(sqrt(n) * (bar(X) - mu) / sigma), ylab = "Tiheys",
     col = "orange", border = "white",
     xlim = c(-4, 4))

# Normaalijakauma kuva 2
x_arvo <- seq(-4, 4, length.out = 100)
y_arvo <- dnorm(x_arvo, mean = 0, sd = 1)
lines(x_arvo, y_arvo, col = "red", lwd = 2)

# Rajat kuva 2
abline(v = c(-1, 1), col = "black", lwd = 2, lty = 2)

legend("topright", legend = c("Z-arvojen välit", "Normaalijakauma"),
       col = c("black", "red"), lty = c(2, 1), lwd = 2)
```

### A.1.5 ARMA(1,1)-KRL

```
# Funktio ARMA(1,1)-funktiolle

generointi_arma11 <- function(n, phi1, theta1, sigma = 1, mu = 0) {
  X <- numeric(n)
  Z <- rnorm(n, mean = 0, sd = sigma)

  # Alustetaan ensimmäinen arvo
  X[1] <- mu +Z[1]

  # Loppujen arvojen generointi
  for (t in 2:n) {
    X[t] <- mu +phi1 * X[t-1] + Z[t] + theta1 * Z[t-1]
```

```
  }

  return(X)
}



set.seed(2025)

# Simulaation parametrit

n <- 100
m <- 10000
phi1 <- 0.2
theta1 <- 0.7
mu <- 0
sigma <- 1

# Stationaarisuuden tarkistus
if (abs(phi1) >= 1) {
  stop("Prosessi ei ole stationaarinen, koska |phi1| >= 1")
}

# sqrt(n)(X_hat-mu) jokaiselle simulaatiolle
simulaatio <- numeric(n)
for (i in 1:m) {
  X <- generointi_arma11(n, phi1, theta1, sigma = sigma, mu =mu)
  X_hat <- mean(X)
  simulaatio[i] <- sqrt(n) * (X_hat - mu)

}



# Histogrammi
hist(simulaatio, breaks = 50, probability = TRUE,
     main = paste("ARMA(1,1), kun phi1 =",phi1,"Theta1 =",theta1,"n =", n),
     xlab = "Arvo",
     ylab = "Tiheys",
     col = "orange")

# Teoreettinen normaalijakauma
X_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
v2 <- sigma^2*((1 + theta1)/(1 - phi1))^2
y_arvo <- dnorm(X_arvo, mean = 0, sd = sqrt(v2))

lines(X_arvo, y_arvo, col = "red", lwd = 2)

legend("topright", legend = c("Simuloitu", "Teoreettinen"),
       col = c("orange", "red"), lwd = c(10,2))



# Tässä on osiossa 9.2 esitetyn todistuksen varmentaminen ja vertailu yllä olevaan
# asymptoottiseen varianssiin (v2)

# Histogrammi
hist(simulaatio, breaks = 50, probability = TRUE,
     main = paste("ARMA(1,1) todistus, kun phi1 =",phi1,"Theta1 =",theta1,"n =", n),
     xlab = "Arvo",
     ylab = "Tiheys",
     col = "blue")


# Sulkujen sisäinen termi ARMA(1,1)-todistus

sisätermi <- (1 + theta1^2 + 2 * phi1 * theta1) +
              (2 * (phi1 + theta1) * (1 + phi1 * theta1)) / (1 - phi1)

# Lopullinen lauseke ARMA(1,1)-todistus
asymp_var <- (sigma^2 / (1 - phi1^2)) * sisätermi
```

```r
y_arvo <- dnorm(X_arvo, mean = 0, sd = sqrt(asymp_var))

lines(X_arvo, y_arvo, col = "red", lwd = 2)

legend("topright", legend = c("Simuloitu", "Teoreettinen"),
       col = c("blue", "red"), lwd = c(10,2))

set.seed(10)

# Yksi ARMA(1,1) sarja
x_sample <-  generointi_arma11(n, phi1, theta1, sigma = sigma, mu =mu)

# Aikasarja
plot(1:n, x_sample, type = "l", col = "orange",
     main = paste("ARMA(1,1)-prosessi ajassa, phi1 =", phi1, ", theta1 =", theta1, "ja n=", n),
     xlab = "Aika", ylab = "X_t")
```

### A.1.6 Vuosittainen aikasarja trendillä (2020 - 2025)

```r
# Havintojen määrä
n <- 16 *12
aika <- 1:n

# Trendin ja kohinan luominen
trendi <- 0.1 * aika
satunnaisuus <- rnorm(n, mean = 0, sd = 3)
X_t <- 50 + trendi + satunnaisuus

# Kuukausittainen aikasraja
aikasarja <- ts(X_t, start = c(2010, 1), frequency = 12)

plot(aikasarja, col = "orange", lwd = 2,
     main = "Vuosittainen aikasraja trendillä (2020 - 2025)",
     ylab = "X_t", xlab = "Vuosi")
grid()
```

## A.2 Koodit simulaatio-osuus

### A.2.1 AR(2)-KRL

```r
    # Laskee hajonnan siten, että prosessin stationaarinen kokonaisvarianssi on target_var
sigma_for_var <- function(phi1, phi2, target_var){
  yule_walker <- (1 - phi2) / ((1 - phi1^2 - phi2^2) * (1 - phi2) - 2 * phi1^2 * phi2)
  sigma2 <- target_var * yule_walker
  return(sqrt(sigma2))
}

# Funktio AR(2)-prosessin generointiin
generointi_ar2 <- function(n, phi1, phi2, target_var= 1, mu = 0) {
  sigma <- sigma_for_var(phi1, phi2, target_var)
  x <- numeric(n)
  # Alkuarvot stationaarisella varianssilla
  x[1] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  x[2] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  for (t in 3:n) {
    x[t] <- mu + phi1 * (x[t - 1] - mu) + phi2 * (x[t - 2] - mu) +
      rnorm(1, mean = 0, sd = sigma)
  }
  return(x)
}

# Simulaatio
set.seed(10)
m <- 10000
phi1 <- 0.2
phi2 <- -0.3
```

```r
mu <- 0
n_arvot <- c(50, 100, 200, 500, 1000, 3000)
target_vars <- c(1, 5, 10, 15)

# Funktio peittotodennäköisyyksien laskentaan
peitto_fun <- function(target_var) {
  sigma <- sigma_for_var(phi1, phi2, target_var)
  asymp_sd <- sigma / abs(1 - phi1 - phi2)

  peitto <- numeric(length(n_arvot))
  for (j in seq_along(n_arvot)) {
    n <- n_arvot[j]
    sisalla <- logical(m)

    for (i in 1:m) {
      x <- generointi_ar2(n, phi1, phi2, target_var = target_var, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper)


    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}

# Peittotodennäköisyydet kaikille variansseille
peitot <- lapply(target_vars, peitto_fun)

plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90, 1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = paste("95% peittotodennäköisyydet AR(2)-KRL, kun phi1 =", phi1, "ja phi2 =", phi2))
for (k in 2:length(target_vars)) {
  lines(n_arvot, peitot[[k]], type = "b", pch = 19, col = k)
}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = paste0("VAR=", target_vars),
       col = seq_along(target_vars), pch = 19, lty = 1)


# Peittotodennäköisyydet näkyvät arvoilla 1-3 ja ne kuvaavat annettuja variansseja
# järjestyksessä. Esim var 1-3 on target_vars <- c(1, 5, 10)

print(data.frame(n = n_arvot,
                 VAR1 = peitot[[1]],
                 VAR2 = peitot[[2]],
                 VAR3 = peitot[[3]],
                 VAR4 = peitot[[3]]))

# Histogrammi jokaiselle varianssille ja n:lle
plot_histogrammit <- function(target_var){
  sigma <- sigma_for_var(phi1, phi2, target_var)
  asymp_sd <- sigma/ abs(1 - phi1 - phi2)

  par(mfrow = c(2,3))
  for(n in n_arvot){
    simulaatio <- numeric(m)
    for(i in 1:m) {
      x <- generointi_ar2(n, phi1, phi2, target_var = target_var, mu = mu)
      x_hat <- mean(x)
      simulaatio[i] <- sqrt(n) * (x_hat - mu)
    }

    hist(simulaatio, breaks = 40, probability = TRUE,
         main = paste("VAR = ", target_var, ", n = ", n),
```

```
        xlab = "Arvo", col = "orange", border = "white")

    # Teoreettinen normaalijakauma
    x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
    y_arvo <- dnorm(x_arvo, mean = 0, sd = asymp_sd)
    lines(x_arvo, y_arvo, col = "red", lwd = 2)
  }
  par(mfrow = c(1,1))
}

for(v in target_vars){
  plot_histogrammit(v)
}

# Simulaatio, kun tutkitaan eri parametrien käyttäytymistä samalla varianssilla

# Laskee hajonnan siten, että prosessin stationaarinen kokonaisvarianssi on target_var
sigma_for_var <- function(phi1, phi2, target_var){
  yule_walker <- (1 - phi2) / ((1 - phi1^2 - phi2^2) * (1 - phi2) - 2 * phi1^2 * phi2)
  sigma2 <- target_var * yule_walker
  return(sqrt(sigma2))
}
# Funktio AR(2)-prosessin generointiin
generointi_ar2 <- function(n, phi1, phi2, target_var= 1, mu = 0) {
  sigma <- sigma_for_var(phi1, phi2, target_var)
  x <- numeric(n)
  # Alkuarvot stationaarisella varianssilla
  x[1] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  x[2] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  for (t in 3:n) {
    x[t] <- mu + phi1 * (x[t - 1] - mu) + phi2 * (x[t - 2] - mu) +
      rnorm(1, mean = 0, sd = sigma)
  }
  return(x)
}

set.seed(10)
m <- 10000

phi1_list <- c(0.7, 0.6, 0.5, 0.4, 0.3)
phi2_list <- c(0.2, 0.2, 0.2, 0.2, 0.2)
mu <- 0
n_arvot <- c(50, 100, 150, 200, 250, 300, 400, 450, 500, 1000)
target_var <- 5

# Funktio peittotodennäköisyyksien laskentaan
peitto_fun <- function(phi1, phi2, target_var) {
  sigma <- sigma_for_var(phi1, phi2, target_var)
  asymp_sd <- sigma / abs(1 - phi1 - phi2)

  peitto <- numeric(length(n_arvot))
  for (j in seq_along(n_arvot)) {
    n <- n_arvot[j]
    sisalla <- logical(m)

    for (i in 1:m) {
      x <- generointi_ar2(n, phi1, phi2, target_var = target_var, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper)


    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}
```

```r
# Simulointi eri phi-arvoilla
peitot <- list()
labels <- character()

for (i in seq_along(phi1_list)) {
  phi1 <- phi1_list[i]
  phi2 <- phi2_list[i]

  peitot[[i]] <- peitto_fun(phi1, phi2, target_var)
  labels[i] <- paste0("phi1 =", phi1, "phi2 =", phi2)

}


plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90,1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = paste("Peittotodennäköisyydet eri parametrien arvoilla AR(2)"))


for (k in 2:length(peitot)) {
  lines(n_arvot, peitot[[k]], type = "b", pch=19, col = k)

}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = labels,
       col = seq_along(labels), pch = 19, lty = 1)

# Histogrammi jokaiselle varianssille ja n:lle
plot_histogrammit <- function(phi1, phi2, target_var){
  sigma <- sigma_for_var(phi1, phi2, target_var)
  asymp_sd <- sigma/ abs(1 - phi1 - phi2)

  par(mfrow = c(2,3))
  for(n in n_arvot){
    simulaatio <- numeric(m)
    for(i in 1:m) {
      x <- generointi_ar2(n, phi1, phi2, target_var = target_var, mu = mu)
      x_hat <- mean(x)
      simulaatio[i] <- sqrt(n) * (x_hat - mu)
    }

    hist(simulaatio, breaks = 40, probability = TRUE,
         main = paste("phi_1 =", phi1, "phi_2 =", phi2, " n = ", n),
         xlab = "Arvo", ylab = "Tiheys", col = "orange", border = "white")

    # Teoreettinen normaalijakauma
    x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
    y_arvo <- dnorm(x_arvo, mean = 0, sd = asymp_sd)
    lines(x_arvo, y_arvo, col = "red", lwd = 2)
  }
  par(mfrow = c(1,1))
}

for(i in seq_along(phi1_list)){
  phi1 <- phi1_list[i]
  phi2 <- phi2_list[i]
  cat("phi_1 =", phi1, "phi_2 =", phi2, "\n")
  plot_histogrammit(phi1, phi2, target_var)

}
```

### A.2.2 MA(2)-KRL

```r
# Funktio MA(2)-prosessin generointiin
generointi_ma2 <- function(n, theta1, theta2, sigma = 1, mu = 0){
  z <- rnorm(n + 2, mean = 0, sd = sigma)
  x <- mu + z[3:(n+2)] + theta1 * z[2:(n+1)] + theta2 * z[1:n]
  return(x)
```

```
}

# Asymptoottinen varianssi
laske_varianssi_ma2 <- function(theta1, theta2, sigma){
  v <- sigma^2 * (1 + theta1 + theta2)^2
  return(v)
}

# Simulaatio
# Simulaation parametrit
#set.seed(10)
m <- 10000
theta1 <- 0.2
theta2 <- 0.7
mu <- 0
n_arvot <- c(50, 100, 200, 500, 1000, 3000)
target_vars <- c(1 ,5, 10, 15)

# Funktio peittotodennäköisyyksien laskemiseen
peitto_fun <- function(target_var){
  sigma <- sqrt(target_var)
  asymp_sd <- sqrt(laske_varianssi_ma2(theta1, theta2, sigma))

  peitto <- numeric(length(n_arvot))
  for(j in seq_along(n_arvot)){
    n <- n_arvot[j]
    sisalla <- logical(m)

    for(i in 1:m){
      x <- generointi_ma2(n, theta1, theta2, sigma = sigma, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper )
    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}

# Peittotodennäköisyydet
peitot <- lapply(target_vars, peitto_fun)

plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90, 1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = "95% peittotodennäköisyydet MA(2)")

for(k in 2:length(target_vars)){
  lines(n_arvot, peitot[[k]], type = "b", pch = 19, col = k)
}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = paste0("VAR", target_vars),
       col = seq_along(target_vars), pch = 19, lty = 1)

# Peittotodennäköisyydet näkyvät arvoilla 1-3 ja ne kuvaavat annettuja variansseja
# järjestyksessä. Esim var 1-3 on target_vars <- c(1, 5, 10)

print(data.frame(n = n_arvot,
                 VAR1 = peitot[[1]],
                 VAR2 = peitot[[2]],
                 VAR3 = peitot[[3]],
                 VAR4 = peitot[[4]]))

# Histogrammit
plot_histogrammit <- function(target_var){
  sigma <- sqrt(target_var)
  asymp_sd <- sqrt(laske_varianssi_ma2(theta1, theta2, sigma))
```

```
  par(mfrow = c(2,3))
  for(n in n_arvot){
    simulaatio <- numeric(m)
    for(i in 1:m){
      x <- generointi_ma2(n ,theta1, theta2, sigma = sigma, mu = mu)
      x_hat <- mean(x)
      simulaatio[i] <- sqrt(n) * (x_hat - mu)
    }

    hist(simulaatio, breaks = 40, probability = TRUE,
         main = paste("VAR = ", target_var, "n =", n),
         xlab = "Arvo", col = "orange", border = "white")

    x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
    y_arvo <- dnorm(x_arvo, mean = 0, sd = asymp_sd)
    lines(x_arvo, y_arvo, col = "red", lwd = 2)
  }
  par(mfrow = c(1,1))
}

for(v in target_vars){
  plot_histogrammit(v)
}

# Simulaatio, kun tutkitaan eri parametrien käyttäytymistä samalla varianssilla

# Funktio MA(2)-prosessin generointiin
generointi_ma2 <- function(n, theta1, theta2, sigma = 1, mu = 0){
  z <- rnorm(n + 2, mean = 0, sd = sigma)
  x <- mu + z[3:(n+2)] + theta1 * z[2:(n+1)] + theta2 * z[1:n]
  return(x)
}

# Asymptoottinen varianssi
laske_varianssi_ma2 <- function(theta1, theta2, sigma){
  v <- sigma^2 * (1 + theta1 + theta2)^2
  return(v)
}

# Simulaation parametrit
#set.seed(10)
m <- 10000
theta1_list <-  c(0.7, 0.6, 0.5, 0.4, 0.3)
theta2_list <- c(0.2, 0.2, 0.2, 0.2, 0.2)
mu <- 0
n_arvot <- c(50, 100, 150, 200, 250, 300, 400, 450, 500, 1000)
target_var <- 5

# Funktio peittotodennäköisyyksien laskemiseen
peitto_fun <- function(theta1, theta2, target_var){
  sigma <- sqrt(target_var)
  asymp_sd <- sqrt(laske_varianssi_ma2(theta1, theta2, sigma))

  peitto <- numeric(length(n_arvot))
  for(j in seq_along(n_arvot)){
    n <- n_arvot[j]
    sisalla <- logical(m)

    for(i in 1:m){
      x <- generointi_ma2(n, theta1, theta2, sigma = sigma, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper )
    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
```

```
 }

 # Simulointi eri phi-arvoilla
peitot <- list()
labels <- character()

for (i in seq_along(theta1_list)) {
  theta1 <- theta1_list[i]
  theta2 <- theta2_list[i]

  peitot[[i]] <- peitto_fun(theta1, theta2, target_var)
  labels[i] <- paste0("theta1 =", theta1, "theta2 =", theta2)

}


plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90,1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = paste("Peittotodennäköisyydet eri parametrien arvoilla MA(2)"))


for (k in 2:length(peitot)) {
  lines(n_arvot, peitot[[k]], type = "b", pch=19, col = k)

}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = labels,
       col = seq_along(labels), pch = 19, lty = 1)
```

### A.2.3 ARMA(1,1)-KRL

```
# Funktio ARMA(1,1)

generointi_arma11 <- function(n, phi1, theta1, sigma = 1, mu = 0) {
  X <- numeric(n)
  Z <- rnorm(n, mean = 0, sd = sigma)

  # Alustetaan ensimmäinen arvo
  X[1] <- mu +Z[1]

  # Loppujen arvojen generointi
  for (t in 2:n) {
    X[t] <- mu +phi1 * X[t-1] + Z[t] + theta1 * Z[t-1]

  }

  return(X)
}

# Prosessin stationaarinen kokonaisvarianssi
arma11_var <- function(phi1, theta1, sigma){
  num <- (1 + theta1^2 +2 * phi1 * theta1) * sigma^2
  den <- (1 - phi1^2)
  return( num / den)
}

# Sigma tietylle varianssille
sigma_for_var <- function(phi1, theta1, target_var){
  factor <- (1 + theta1^2 +2 * phi1 * theta1) / (1 - phi1^2)
  sigma2 <- target_var/ factor
  return(sqrt(sigma2))
}

# Asymptoottisen varianssin laskeminen
laske_varianssi <- function(phi1, theta1, sigma) {
  varianssi <- sigma^2 * ((1 + theta1) / (1 - phi1))^2
  return(varianssi)
}
```

```r
# Simulaation parametrit
#set.seed(10)
m <- 10000
phi1 <- 0.2
theta1 <- 0.7
mu <- 0
n_arvot <- c(50 , 100, 200, 500, 1000, 3000)
target_vars <- c(1, 5, 10, 15)

# Funktio peittotodennäköisyyskien laskemiseen
peitto_fun <- function(target_var){
  sigma <- sigma_for_var(phi1, theta1, target_var)
  asymp_var <- laske_varianssi(phi1, theta1, sigma)
  asymp_sd <- sqrt(asymp_var)

  peitto <- numeric(length(n_arvot))
  for (j in seq_along(n_arvot)){
    n <- n_arvot[j]
    sisalla <- logical(m)

    for (i in 1:m) {
      x <- generointi_arma11(n, phi1, theta1, sigma = sigma, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper)
    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}

# Peitto todennäköisyys kaikille
peitot <- lapply(target_vars, peitto_fun)

plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90,1), xlab = "n", ylab = "Peittotodennakoisyys",
     main = "95% peittotodennakoisyydet ARMA(1,1)")
for(k in 2:length(target_vars)){
  lines(n_arvot, peitot[[k]], type = "b", pch = 19, col = k)
}

abline(h = 0.95, col = "red", lty = 2, lwd =2)
legend("bottomright", legend = paste0("VAR=", target_vars),
       col = seq_along(target_vars), pch = 19, lty = 1)
```
```r
# Peittotodennäköisyydet näkyvät arvoilla 1-3 ja ne kuvaavat annettuja variansseja
# järjestyksessä. Esim var 1-3 on target_vars <- c(1, 5, 10)

print(data.frame(n = n_arvot,
                 VAR1 = peitot[[1]],
                 VAR2 = peitot[[2]],
                 VAR3 = peitot[[3]],
                 VAR4 = peitot[[4]]))
# Histogrammi jokaiselle varianssille ja n:lle
plot_histogrammit <- function(target_var){
  sigma <- sqrt(target_var)
  asymp_sd <- sqrt(laske_varianssi(phi1, theta1, sigma))

  par(mfrow = c(2,3))
  for(n in n_arvot){
    simulaatio <- numeric(m)
    for(i in 1:m) {
      x <- generointi_arma11(n, phi1, theta1, sigma = sigma, mu = mu)
      x_hat <- mean(x)
      simulaatio[i] <- sqrt(n) * (x_hat - mu)
    }
```

```
    hist(simulaatio, breaks = 40, probability = TRUE,
         main = paste("VAR = ", target_var, ", n = ", n),
         xlab = "Arvo", col = "orange", border = "white")

    x_arvo <- seq(min(simulaatio), max(simulaatio), length.out = 500)
    y_arvo <- dnorm(x_arvo, mean = 0, sd = asymp_sd)
    lines(x_arvo, y_arvo, col = "red", lwd = 2)
  }
  par(mfrow = c(1,1))
}

for(v in target_vars){
  plot_histogrammit(v)
}

# Simulaatio, kun tutkitaan eri parametrien käyttäytymistä samalla varianssilla

# Funktio ARMA(1,1)

generointi_arma11 <- function(n, phi1, theta1, sigma = 1, mu = 0) {
  X <- numeric(n)
  Z <- rnorm(n, mean = 0, sd = sigma)

  # Alustetaan ensimmäinen arvo
  X[1] <- mu +Z[1]

  # Loppujen arvojen generointi
  for (t in 2:n) {
    X[t] <- mu +phi1 * X[t-1] + Z[t] + theta1 * Z[t-1]

  }

  return(X)
}

# Prosessin stationaarinen kokonaisvarianssi
arma11_var <- function(phi1, theta1, sigma){
  num <- (1 + theta1^2 +2 * phi1 * theta1) * sigma^2
  den <- (1 - phi1^2)
  return( num / den)
}

# Sigma tietylle varianssille
sigma_for_var <- function(phi1, theta1, target_var){
  factor <- (1 + theta1^2 +2 * phi1 * theta1) / (1 - phi1^2)
  sigma2 <- target_var/ factor
  return(sqrt(sigma2))
}

# Asymptoottisen varianssin laskeminen
laske_varianssi <- function(phi1, theta1, sigma) {
  varianssi <- sigma^2 * ((1 + theta1) / (1 - phi1))^2
  return(varianssi)
}

# Simulaation parametrit
#set.seed(10)
m <- 10000
phi1_list <- c(0.7, 0.6, 0.5, 0.4, 0.3)
theta1_list <-  c(0.2, 0.2, 0.2, 0.2, 0.2)
mu <- 0
n_arvot <- c(50, 100, 150, 200, 250, 300, 400, 450, 500, 1000)
target_var <- 5

# Funktio peittotodennäköisyyskien laskemiseen
peitto_fun <- function(phi1, theta1, target_var){
  sigma <- sigma_for_var(phi1, theta1, target_var)
  asymp_var <- laske_varianssi(phi1, theta1, sigma)
  asymp_sd <- sqrt(asymp_var)

  peitto <- numeric(length(n_arvot))
```

```
  for (j in seq_along(n_arvot)){
    n <- n_arvot[j]
    sisalla <- logical(m)

    for (i in 1:m) {
      x <- generointi_arma11(n, phi1, theta1, sigma = sigma, mu = mu)
      x_hat <- mean(x)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd / sqrt(n)
      upper <- x_hat + z * asymp_sd / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper)
    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}

# Simulointi eri phi-arvoilla
peitot <- list()
labels <- character()

for (i in seq_along(phi1_list)) {
  phi1 <- phi1_list[i]
  theta1 <- theta1_list[i]

  peitot[[i]] <- peitto_fun(phi1, theta1, target_var)
  labels[i] <- paste0("phi1 =", phi1, "theta1 =", theta1)

}


plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90,1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = paste("Peittotodennäköisyydet eri parametrien arvoilla ARMA(1,1)"))


for (k in 2:length(peitot)) {
  lines(n_arvot, peitot[[k]], type = "b", pch=19, col = k)

}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = labels,
       col = seq_along(labels), pch = 19, lty = 1)
```

### A.2.4 AR(2)-Yule-Walker

```
# Laskee hajonnan siten, että prosessin stationaarinen kokonaisvarianssi on target_var
sigma_for_var <- function(phi1, phi2, target_var){
  yule_walker <- (1 - phi2) / ((1 - phi1^2 - phi2^2) * (1 - phi2) - 2 * phi1^2 * phi2)
  sigma2 <- target_var * yule_walker
  return(sqrt(sigma2))
}

# Funktio AR(2)-prosessin generointiin
generointi_ar2 <- function(n, phi1, phi2, target_var= 1, mu = 0) {
  sigma <- sigma_for_var(phi1, phi2, target_var)
  x <- numeric(n)
  # Alkuarvot stationaarisella varianssilla
  x[1] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  x[2] <- rnorm(1, mean = mu, sd = sqrt(target_var))
  for (t in 3:n) {
    x[t] <- mu + phi1 * (x[t - 1] - mu) + phi2 * (x[t - 2] - mu) +
      rnorm(1, mean = 0, sd = sigma)
  }
  return(x)
}
```

```r
# Simulaatio
set.seed(10)
m <- 10000
phi1 <- 0.2
phi2 <-  0.7
mu <- 0
n_arvot <- c(50, 100, 200, 500, 1000, 3000)
target_vars <- c(1, 5, 10, 15)

# Funktio peittotodennäköisyyksien laskentaan
peitto_fun <- function(target_var) {
  peitto <- numeric(length(n_arvot))
  for (j in seq_along(n_arvot)) {
    n <- n_arvot[j]
    sisalla <- logical(m)

    for (i in 1:m) {
      x <- generointi_ar2(n, phi1, phi2, target_var = target_var, mu = mu)
      x_hat <- mean(x)


      # AR(2) estimointi Walker
      ar_sov <- ar.yw(x, order.max = 2, aic = FALSE)
      phi1_hat <- ar_sov$ar[1]
      phi2_hat <- ar_sov$ar[2]
      sigma_hat <- sqrt(ar_sov$var.pred)


      # Estimointipohjainen asymptoottinen hajonta Walker
      asymp_sd_hat <- sigma_hat / abs(1 - phi1_hat - phi2_hat)

      # 95% luottamusväli
      z <- qnorm(0.975)
      lower <- x_hat - z * asymp_sd_hat / sqrt(n)
      upper <- x_hat + z * asymp_sd_hat / sqrt(n)
      sisalla[i] <- (mu >= lower & mu <= upper)


    }
    peitto[j] <- mean(sisalla)
  }
  return(peitto)
}

# Peittotodennäköisyydet kaikille variansseille
peitot <- lapply(target_vars, peitto_fun)

plot(n_arvot, peitot[[1]], type = "b", pch = 19, col = 1,
     ylim = c(0.90, 1), xlab = "n", ylab = "Peittotodennäköisyys",
     main = paste("Peittotodennäköisyydet AR(2)-Walker, kun phi1 =", phi1, "ja phi2 =", phi2))

for (k in 2:length(target_vars)) {
  lines(n_arvot, peitot[[k]], type = "b", pch = 19, col = k)
}

abline(h = 0.95, col = "red", lty = 2, lwd = 2)

legend("bottomright", legend = paste0("VAR=", target_vars),
       col = seq_along(target_vars), pch = 19, lty = 1)
```