TAMPEREEN YLIOPISTO
Pro gradu -tutkielma

Jari-Pekka Piiroinen

Ajoneuvovakuutusten hinnoittelu tilastollisista lähtökohdista

Luonnontieteiden tiedekunta
Matematiikka
Toukokuu 2017

Tampereen yliopisto

Luonnontieteiden tiedekunta

Piiroinen, Jari-Pekka: Ajoneuvovakuutusten hinnoittelu tilastollisista lähtökohdista

Pro gradu -tutkielma, 47 s., 6 liites.

Matematiikka

Toukokuu 2017

Tiivistelmä

Tutkielmassa tutustutaan ajoneuvovakuutusten hinnoitteluun yleistettyjen lineaaristen mallien avulla. Yleistettyjen lineaaristen mallien teoriaan liittyen tutustumme eksponentiaaliseen jakaumaperheeseen ja erityisesti tweedie-jakaumaperheeseen. Tweedie-jakaumaperhe sisältää erikoistapauksena Poisson-gamma-jakauman, joka on samalla diskreetti että jatkuva jakauma. Poisson-gamma-jakauman avulla pystymme yhdistämään vahinkotiheyden ja keskivahingon mallintamisen erikseen mallintamisen sijasta. Itse yleistettyjen lineaaristen mallien teoriaa käymme kursorisesti läpi ja keskitymme hinnoittelumallin rakentamiseen ja eri tariffitekijöiden luokkien suhteellisten riskien estimointiin. Aineiston soveltavaa osuutta varten olemme saaneet Suomen Vahinkovakuutukselta.

Tutkielman matemaattinen osuus koostuu mitta- ja todennäköisyysteoriasta. Mittateoriassa tutustutaan mitan ja mitallisen kuvauksen määritelmään sekä niiden ominaisuuksiin. Mittateorian tuloksista tärkeimpänä esittelemme Caratheodoryn laajennuslauseen, jonka avulla konstruoimme Lebesgue-Stieltjesin- ja Lebesguen mitan. Mittateorian osuudella motivoimme työssä käsiteltävien todennäköisyysteorian käsitteitä. Todennäköisyysteoria voidaan nähdä mittateorian erityistapauksena, jolla luodaan perusta tilastolliselle päättelylle. Esittelemme työlle tärkeät satunnaismuuttujan, todennäköisyysjakauman ja odotusarvon käsitteet. Lisäksi kiinnitämme huomiota momentteihin ja momentit generoiviin funktioihin, joiden avulla voidaan luonnehtia jakaumien ominaisuuksia. Määrittelemme erityisesti momentit generoivien funktioiden avulla eksponentiaaliseen jakaumaperheeseen kuuluvien jakaumien odotusarvot ja varianssit.

Sisältö

1 Johdanto

Tämä työ käsittelee yleistettyjä lineaarisia malleja vakuutusmatemaattisesta näkökulmasta ja siitä, miten yleistettyjä lineaarisia malleja käytetään vahinkovakuutusyhtiöissä tuotteiden hinnoitteluun. Yleistetyt lineaariset mallit ovat laajasti käytössä vahinkovakuutusyhtiöiden toiminnassa ja on suhteellisen uusi menetelmä tilastotieteessä ja vakuutustuotteiden hinnoittelussa. Varsinainen yleistettyjen lineaaristen mallien teoria kehitettiin vuonna 1972 John Nelderin ja Robert Wedderburnin toimesta. Vakuutustuotteiden hinnoittelussa yleistettyjä lineaarisia malleja on käytetty 1980-luvulta lähtien.

Perinteisesti vakuutustuotteiden hinnoittelussa on käytetty multiplikatiivista mallia käyttäen log-linkkifunktiota, jolloin eri tariffiluokkien riskikertoimet ovat suhteellisia eri tariffiluokkien kesken. Hinnoittelu tapahtuu mallintamalla vahinkotiheyttä käyttäen Poisson-jakauman oletuksia, ja mallintamalla vahinkojen suuruutta käyttäen Gamma-jakaumaa. Esimerkiksi kirjassa *Non-Life Insurance Pricing with Generalized Linear Models* [2] paneudutaan hinnoittelukysymyksiin yleistettyjen lineaaristen mallien näkökulmasta. Kuitenkin tässä tutkielmassa käytämme yhdistettyä Poisson-gamma-jakaumaa, joka kuuluu tweedie-jakaumaperheeseen. Yhdistetty Poisson-gamma-jakauma yhdistää jatkuvan ja diskreetin jakauman ominaisuudet, jolloin estimointi tapahtuu kerralla eikä vahinkojen suuruuden ja vahinkotiheyden estimoinnin tuloksia tarvitse yhdistää. Vahinkotiheyden ja vahinkojen suuruuden tuloa kutsumme puhtaaksi preemioksi. Tässä tutkielmassa käytetään Suomen Vahinkovakuutukselta saatua vahinkoaineistoa, johon sovelletaan tutkielmassa esitettyjä menetelmiä. Tulosten tulkintaan käytettävää teoriaa ei esitellä, mutta annetaan viite mahdollisia jatkotarkasteluja varten.

Luvussa 2 käsitellään ensin mittateorian perusteita, jossa käymme läpi mitan ja mitallisten kuvausten määritelmän sekä niiden ominaisuuksia. Esitämme myös tärkeitä mittateoreettisia tuloksia, joista tärkeimpänä Caratheodoryn laajennuslause. Caratheodoryn laajennuslauseen avulla konstruoimme alaluvussa 2.4 Lebesgue-Stieltjesin-mitan ja Lebesguen mitan, jotka ovat käytännöllisiä todennäköisyysteoriassa. Luvussa 3 käsittelemme todennäköisyysteoriaa, ja määrittelemme satunnaismuuttujan ja todennäköisyysjakauman käsitteet. Alaluvussa 3.2 käsittelemme odotusarvon käsitettä ja luvussa 3.3 käymme kursorisesti läpi momenttien luonnetta ja niiden yhteyttä jakauman odotusarvoon ja varianssiin. Varsinaisen mitta- ja todennäköisyysteorian osuuden jälkeen käsittelemme luvussa 4 tilastollisia esitietoja, joista tärkeimpinä eksponentiaaliset hajontaperheet alaluvussa 4.1 ja tämän osajoukko Tweedie-jakaumaperhe alaluvussa 4.2. Itse soveltava osuus on luvussa 5, jossa sovellamme yleistettyjä lineaarisia malleja vakuutusmatemaattiseen ongelmaan. Soveltavan osuuden tarkoituksena on esitellä teorian yhteyttä käytännön ongelmaan, jossa tarkastelemme Suomen Vahinkovakuutuksen pakettiautojen vahinkotilastoja ja estimoimme liikennevakuutuksen puhdasta preemiota käyttäen Tweedie-jakaumaperhettä.

Työ itsessään on hyvin mittateoreettinen, ja lähtökohtana yleistettyjen lineaaristen malleihin on todennäköisyysteoreettinen näkökulma, jonka pohjalle tämä tutkielma perustuu. Lukijalta odotetaan analyysin tuntemusta ja joukko-opin perusteiden hallintaa.

2 Mittateoriaa

Tässä luvussa esitämme mittateorian perusteita. Mittateoriaa tarvitsemme luvussa 3 todennäköisyysteoriaa varten, jota sovellamme luvun 4 tilastolliseen osuuteen. Mittateorian osuus perustuu kirjoihin *Probability with Martingales* [1, Williams], *Real Analysis and Probability* [4, Ash] ja *Real and Complex Analysis* [5, Rudin], joista voi halutessaan tarkastella puuttuvia todistuksia ja täydentää omaa tietämystään. Huomaa, että tutkielmassa luonnollisten lukujen joukko $\mathbb{N}$ on määritelty joukkona $\mathbb{N} = \{1, 2, 3, \dots\}$.

2.1 Mittateorian perusteita

Määritelmä 2.1. Olkoon Ω joukko ja $\mathcal{P}(\Omega)$ joukon Ω potenssijoukko. Olkoon $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ kokoelma joukon Ω osajoukkoja. Kokoelmaa $\mathcal{A}$ sanotaan *algebraksi*, kun

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c = (\Omega \backslash A) \in \mathcal{A}$
3. $A, B \in \mathcal{A} \Rightarrow (A \cup B) \in \mathcal{A}$

Lisäksi koska $A \cap B = (A^c \cup B^c)^c$, niin määritelmästä seuraa, että algebra on suljettu äärellisten leikkausten ja yhdisteiden suhteen.

Määritelmä 2.2. Algebra $\mathcal{A}$ on *σ-algebra*, kun se on suljettu numeroituvan yhdisteen suhteen, eli

$$\{A_n \colon n \in \mathbb{N}\} \subseteq \mathcal{A} \Rightarrow \left(\bigcup_{n=1}^{\infty} A_n \right) \in \mathcal{A}.$$

Määritelmästä seuraa, että σ-algebra on suljettu myös numeroituvan leikkauksen suhteen. Tällöin kutsumme paria $(\Omega, \mathcal{A})$ *mitalliseksi avaruudeksi*. Lisäksi σ-algebran jäseniä $A \in \mathcal{A}$ kutsutaan *mitallisiksi joukoiksi*.

Esimerkki 2.1. Olkoon Ω joukko.

(i) Kokoelma $\{\emptyset, \Omega\}$ on joukon Ω suppein σ-algebra.

(ii) Potenssijoukko $\mathcal{P}(\Omega)$ on joukon Ω laajin σ-algebra.

(iii) Kun $A \subset \Omega$, niin kokoelma $\{\emptyset, A, A^c, \Omega\}$ on σ-algebra.

(iv) Kun $\Omega = \mathbb{N}$, niin kokoelma $\{\emptyset, \{1, 3, 5, \dots\}, \{2, 4, 6, \dots\}, \Omega\}$ on *sigma*-algebra.

Lause 2.1. *Olkoon $\mathcal{C}$ kokoelma joukon Ω osajoukkoja. Tällöin on olemassa joukon Ω pienin σ-algebra $\sigma(\mathcal{C})$, jolle $\mathcal{C} \subseteq \sigma(\mathcal{C})$.*

Todistus. [5, s. 12] Olkoon $\mathcal{G}$ niiden σ-algebrojen $\mathcal{A}$ kokoelma joukossa Ω, joihin $\mathcal{C}$ sisältyy. Koska joukon Ω potenssijoukko $\mathcal{P}(\Omega) \in \mathcal{G}$, niin kokoelma $\mathcal{G}$ on epätyhjä.

Olkoon $\sigma(\mathcal{C})$ leikkaus kaikista joukoista $\mathcal{A}$, joilla $\mathcal{A} \in \mathcal{G}$. On selvää, että $\mathcal{C} \subseteq \sigma(\mathcal{C})$ ja $\sigma(\mathcal{C})$ sisältyy jokaiseen σ-algebraan joukossa Ω, jotka sisältävät joukon $\mathcal{C}$.

Osoitetaan vielä, että $\sigma(\mathcal{C})$ on σ-algebra. Jos $A_n \in \sigma(\mathcal{C})$ kaikilla $n \in \mathbb{N}$, ja jos $\mathcal{A} \in \mathcal{G}$, niin $\cup A_n \in \mathcal{A}$, sillä $\mathcal{A}$ on σ-algebra. Koska $\cup A_n \in \mathcal{A}$ jokaisella $\mathcal{A} \in \mathcal{G}$, niin myös $\cup A_n \in \sigma(\mathcal{C})$. Kaksi muuta σ-algebran ominaisuutta voidaan todistaa vastaavasti. □

Määritelmä 2.3. Olkoon $(S, \mathcal{T})$ topologinen avaruus, jossa topologia $\mathcal{T}$ koostuu kaikista S:n avoimista joukoista. *Borelin σ-algebra* $\mathcal{B}(S) := \sigma(\mathcal{T})$ on avointen joukkojen virittämä σ-algebra.

Borelin σ-algebran $\mathcal{B}(S)$ jäseniä kutsutaan Borelin joukoiksi. Yleinen Borelin joukko voi olla liian monimutkainen esitettäväksi, joten on hyödyllistä tietää yksinkertaisempi joukkoperhe, joka virittää Borelin σ-algebran numeroituvien yhdisteiden ja leikkausten kautta.

Esimerkki 2.2. Vrt. [1, s. 17] Olkoon $(S, \mathcal{T})$ topologinen avaruus, $S = \mathbb{R}$ ja $\mathcal{T}$ reaalilukujen tavallinen topologia. Silloin

$$\mathcal{A} := \sigma\{(-\infty, x] \colon x \in \mathbb{R}\} = \sigma(\mathcal{T}) = \mathcal{B}(\mathbb{R}).$$

Todistus. Koska

$$(-\infty, x] = \bigcap_{n \in \mathbb{N}} (-\infty, x + n^{-1})$$

on Borelin joukko, niin siitä seuraa, että $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R})$. Kun $a, b \in \mathbb{R}$ ja $a < b$, niin

$$(a, b) = (-\infty, a]^c \cap \left(\bigcup_{n \in \mathbb{N}} (-\infty, b - n^{-1}] \right) \in \mathcal{A}.$$

Olkoon $U \subseteq \mathbb{R}$ avoin. Jokaiselle $q \in U \cap \mathbb{Q}$ on olemassa $\epsilon_q > 0$, jolla $(q - \epsilon_q, q + \epsilon_q) \subseteq U$. Koska $\mathbb{Q}$ on tiheä $\mathbb{R}$:ssä, niin seuraa että

$$U = \bigcup_{q \in U \cap \mathbb{Q}} (q - \epsilon_q, q + \epsilon_q),$$

jossa yhdiste on numeroituva. Tästä seuraa, että $U \in \mathcal{A}$ ja siksi $\mathcal{B}(\mathbb{R}) = \mathcal{A}$. □

Määritelmä 2.4. Olkoon $(\Omega, \mathcal{A})$ mitallinen avaruus. Kutsumme kuvausta $\mu \colon \mathcal{A} \to [0, \infty]$ *positiiviseksi mitaksi*, kun $\mu(\emptyset) = 0$ ja se on *numeroituvasti additiivinen* (σ-additiivinen), eli kaikille jonoille $\{A_n \colon n \in \mathbb{N}\} \subseteq \mathcal{A}$, jossa joukot ovat erilliset, pätee

$$\mu\left(\bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{2.1}$$

Lisäksi mitallinen avaruus on *mitta-avaruus* $(\Omega, \mathcal{A}, \mu)$, kun sen mitallisissa joukoissa on määritelty positiivinen mitta.

Huomautus. Vaikka käsittelemme todennäköisyysteoriaa vasta luvussa 3, niin määrittelemme jo tässä vaiheessa *todennäköisyysmitan* käsitteen alaluvussa 2.2 käsiteltävää *Dynkinin laajennuslausetta* 2.1 varten.

Määritelmä 2.5. *Todennäköisyysmitta* on funktio $P\colon \Omega \to [0,1]$, jolle

(i) $P(\Omega) = 1$,

(ii) P on positiivinen mitta.

Tällöin avaruutta $(\Omega, \mathcal{A}, P)$ kutsutaan *todennäköisyysavaruudeksi*. Joukkoa Ω kutsumme *otosavaruudeksi* ja sen alkioita *alkeistapauksiksi*. Kokoelman $\mathcal{A}$ alkiot ovat *tapahtumia*.

Määritelmä 2.6. Olkoon $(\Omega, \mathcal{A}, \mu)$ mitta-avaruus. Mitan μ sanotaan olevan *äärellinen*, kun $\mu(\Omega) < \infty$ ja *σ-äärellinen*, jos on olemassa numeroituva mitallinen peite $\{\Omega_n\colon n \in \mathbb{N}\}$, jolla pätee

$$\mu(\Omega_n) < \infty \quad \text{ja} \quad \Omega = \bigcup_{i=0}^{\infty} \Omega_i.$$

Määritelmä 2.7. Olkoon P todennäköisyysmitta todennäköisyysavaruudessa $(\Omega, \mathcal{A})$. Kun $A \in \mathcal{A}$ ja $P(A) = 1$, sanotaan että A tapahtuu P-melkein varmasti (m.v.). Vastaavasti joukko A^c on P-nollajoukko.

Määritelmä 2.8. Olkoot $(\Omega, \mathcal{A})$ mitallinen avaruus ja joukot A_n, B_n joukon Ω osajoukkoja kaikilla $n \in \mathbb{N}$.

(i) Jos kaikilla $n \in \mathbb{N}$ pätee $A_n \subseteq A_{n+1}$ ja $\cup A_n = A$, niin merkitään $A_n \uparrow A$.

(ii) Jos kaikilla $n \in \mathbb{N}$ pätee $B_{n+1} \subseteq B_n$ ja $\cap B_n = B$, niin merkintään $B_n \downarrow B$.

Lause 2.2 (Mitan monotoninen suppeneminen)**.** *Olkoon $(\Omega, \mathcal{A}, \mu)$ mitta-avaruus.*

(i) Jos $A_n \in \mathcal{A}$ kaikilla $n \in \mathbb{N}$ ja $A_n \uparrow A$, niin silloin $\lim_{n\to\infty} \mu(A_n) = \mu\left(\bigcup_{n=1}^{\infty} A_n\right)$. Tällöin merkitsemme $\mu(A_n) \uparrow \mu(A)$.

(ii) Jos $B_n \in \mathcal{A}$ kaikilla $n \in \mathbb{N}$, $B_n \downarrow B$ ja $\mu(B_n) < \infty$ jollakin $n \in \mathbb{N}$, niin silloin $\lim_{n\to\infty} \mu(B_n) = \mu\left(\bigcap_{n=1}^{\infty} B_n\right)$. Tällöin merkitsemme $\mu(B_n) \downarrow \mu(B)$.

Todistus. Vrt. [1, s. 21-22]

(i) Merkitään $C_{n+1} := A_{n+1} \backslash A_n$, kun $n \in \mathbb{N}$. Nyt $A = \bigcup_{n\in\mathbb{N}} C_n$, missä joukot C_n ovat erillisiä. Siis

$$\mu(A) = \sum_{k=1}^{\infty} \mu(C_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(C_k) = \lim_{n\to\infty} \left(\bigcup_{k=1}^{n} \mu(C_k) \right) = \lim_{n\to\infty} \mu(A_n).$$

(ii) Olkoon $A_n := B_1 \backslash B_n$, kun $n \in \mathbb{N}$. Tällöin joukot $A_1 \subset A_2 \subset \ldots$ ovat mitallisia. Todistus menee vastaavasti kuin edellinen kohta.

□

2.2 Mitan laajennuslauseet

Tässä alaluvussa esittelemme Dynkinin ja Carathéodoryn laajennuslauseet. Erityisesti Carathéodoryn laajennuslause on tärkeä mittateoreettinen tulos. Carathéodoryn laajennuslausetta emme todista tässä tutkielmassa, vaan annamme todistukseen viitteen lähempää tarkastelua varten.

Määritelmä 2.9. Joukkoperhettä $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ kutsutaan *Dynkinin luokaksi*, eli *d-systeemiksi*, kun

1. $\Omega \in \mathcal{D}$,
2. $A, B \in \mathcal{D}$ ja $A \subseteq B \Rightarrow B \backslash A \in \mathcal{D}$,
3. $\{A_n \colon n \in \mathbb{N}\} \subseteq \mathcal{D}$ ja $A_n \uparrow A \Rightarrow A \in \mathcal{D}$.

Määritelmä 2.10. Joukkoperhettä $\mathcal{I} \subseteq \mathcal{P}(\Omega)$ kutsutaan *π-systeemiksi*, jos se on suljettu äärellisten leikkausten suhteen, eli

$$I_1, I_2 \in \mathcal{I} \Rightarrow I_1 \cap I_2 \in \mathcal{I}.$$

Lemma 2.1. Olkoon $\mathcal{A}$ avaruuden Ω kokoelma. Tällöin se on σ-algebra, jos ja vain jos se on sekä π-systeemi että d-systeemi.

Todistus. Vrt. [1, s. 193] Selvästi σ-algebra on sekä d- että π-systeemi.

Toisinpäin riittää osoittaa, että numeroituvan yhdisteen ominaisuus on voimassa. Olkoon $\{B_n \colon n \in \mathbb{N}\} \subseteq \mathcal{A}$. Oletuksesta seuraa, että äärelliset yhdisteet kuuluvat joukkoon $\mathcal{A}$, joten

$$A_n := B_1 \cup B_2 \cup \cdots \cup B_n = (B_1^c \cap B_2^c \cap \cdots \cap B_n^c)^c \in \mathcal{A}$$

Lisäksi koska nouseva ketju A_n suppenee kohti joukkoa A, niin

$$A_n \uparrow A := \left(\bigcup_{n\in\mathbb{N}} A_n \right) = \left(\bigcup_{n\in\mathbb{N}} B_n \right) \in \mathcal{A}.$$

Siis $\mathcal{A}$ toteuttaa σ-algebran ominaisuudet. □

Lause 2.3 (Dynkinin π-d-lemma)**.** *Olkoon $\mathcal{I}$ π-systeemi avaruudessa Ω. Tällöin pienin d-systeemi, joka sisältää $\mathcal{I}$:n on σ-algebra, eli $d(\mathcal{I}) = \sigma(\mathcal{I})$.*

Todistus. Vrt. [1, s. 193-194] Edellisen lemman vuoksi riittää osoittaa, että $d(\mathcal{I})$ on π-systeemi. Osoitetaan ensin, että

$$\mathcal{D}_1 := \{B \in d(\mathcal{I}) \,|\, \forall E \in \mathcal{I} \colon B \cap E \in d(\mathcal{I})\},$$

on d-systeemi. Nähdään suoraan, että $\mathcal{I} \subseteq \mathcal{D}_1 \subseteq d(\mathcal{I})$.

(i) Koska $\Omega \in d(\mathcal{I})$ ja $(\Omega \cap E) = E \in \mathcal{I}$ kaikilla $E \in \mathcal{I}$, niin $\Omega \in \mathcal{D}_1$.

(ii) Olkoon $B_1, B_2 \in \mathcal{D}_1$. Jos $B_1 \subseteq B_2$, niin d-systeemin määritelmän ja joukon $\mathcal{D}_1$ ominaisuuksien mukaan $(B_2 \backslash B_1) \in d(\mathcal{I})$, ja kaikilla $E \in \mathcal{I}$ pätee $(B_2 \backslash B_1) \cap E = (B_2 \cap E) \backslash (B_1 \cap E)$. Nyt siis $\mathcal{D}_1$:n määritelmän mukaan $(B_1 \cap E), (B_2 \cap E) \in d(\mathcal{I})$. Koska $d(\mathcal{I})$ on d-luokka, niin seuraa että kaikilla $E \in \mathcal{I}$ pätee $(B_2 \backslash B_1) \cap E \in d(\mathcal{I})$, joten $(B_2 \backslash B_1) \in \mathcal{D}_1$.

(iii) Olkoon $\{B_n \colon n \in \mathbb{N}\} \subseteq \mathcal{D}_1$, joka suppenee kohti joukkoa B, eli $B_n \uparrow B = \cup_n B_n$. Tällöin kaikille $E \in \mathcal{I}$ pätee $B \cap E = \bigcup_n (B_n \cap E)$, jossa $(B_n \cap E)$ suppenee kohti joukkoa $(B \cap E)$. Nyt $\mathcal{D}_1$:n määritelmän mukaan $(B_n \cap E) \in d(\mathcal{I})$. Koska $d(\mathcal{I})$ on d-luokka, niin kaikilla $E \in \mathcal{I}$ pätee $(B \cap E) \in \mathcal{D}_1$.

Nyt on osoitettu, että $\mathcal{D}_1$ on d-luokka, ja koska $\mathcal{I} \subseteq \mathcal{D}_1 \subseteq d(\mathcal{I})$, niin seuraa että $\mathcal{D}_1 = d(\mathcal{I})$.
Olkoon nyt

$$\mathcal{D}_2 := \{B \in d(\mathcal{I}) \,|\, \forall A \in d(\mathcal{I}) \colon B \cap A \in d(\mathcal{I})\} \subseteq d(\mathcal{I}).$$

Olemme osoittaneet, että jos $E \in \mathcal{I}$ ja $B \in \mathcal{D}_1 = d(\mathcal{I})$, niin silloin $(E \cap B) \in d(\mathcal{I})$. Tästä seuraa, että $\mathcal{I} \subseteq \mathcal{D}_2$. Edellisen kohdan perusteella voidaan osoittaa, että $\mathcal{D}_2$ on d-luokka, joten $\mathcal{D}_2 = d(\mathcal{I})$. Tästä seuraa, että d-luokka $d(\mathcal{I})$ on π-luokka, ja lemman 2.1 perusteella $d(\mathcal{I})$ on σ-algebra, yleisesti siis $d(\mathcal{I}) \subseteq \sigma(\mathcal{I})$. Koska $d(\mathcal{I})$ on σ-algebra, joka sisältää $\mathcal{I}$:n, niin seuraa, että $\sigma(\mathcal{I}) \subseteq d(\mathcal{I})$, eli $\sigma(\mathcal{I}) = d(\mathcal{I})$. □

Seuraus 2.1 (Dynkinin laajennuslause)**.** Olkoon $\mathcal{I}$ π-systeemi ja olkoon $\mathcal{A} := \sigma(\mathcal{I})$. Olkoot P ja Q kaksi todennäköisyysmittaa mitallisessa avaruudessa $(\Omega, \mathcal{A})$. Jos $P(I) = Q(I)$ jokaisella $I \in \mathcal{I}$, niin $P(A) = Q(A)$ jokaisella $A \in \mathcal{A}$.

Todistus. Vrt. [1, s. 194-195] Olkoon

$$\mathcal{D} := \{A \in \mathcal{A} \,|\, P(A) = Q(A)\}.$$

Osoitamme, että $\mathcal{D}$ on d-systeemi, jolloin Dynkinin lemman perusteella $\mathcal{D} = \sigma(\mathcal{I}) = \mathcal{A}$. Nähdään, että $\Omega \in \mathcal{D}$. Olkoon $A, B \in \mathcal{D}$ ja $A \subseteq B$. Tällöin

$$P(B\backslash A) = P(B) - P(A) = Q(B) - Q(A) = Q(B\backslash A).$$

Siis $B\backslash A \in \mathcal{D}$. Olkoon $(A)_{n\in\mathbb{N}} \in \mathcal{D}$ nouseva ketju, joka suppenee kohti joukkoa A, eli $A_n \uparrow A$. Nyt

$$P(A) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} Q(A_n) = Q(A),$$

joten $A \in \mathcal{D}$. Siis $\mathcal{D}$ on d-systeemi, jolle $\mathcal{I} \subseteq \mathcal{D}$. Nyt Dynkinin lemman mukaan $\mathcal{A} = \sigma(\mathcal{I}) = d(\mathcal{I}) \subseteq \mathcal{D}$, mikä todistaa väitteen. □

Palautetaan mieleen määritelmä 2.1. Funktiota $\mu\colon \mathcal{A} \to [0,\infty]$ kutsumme tällöin *esimitaksi*, jos se on numeroituvasti additiivinen. Sanomme, että esimitta on *σ-äärellinen*, jos on olemassa joukot $A_i \in \mathcal{A}$, joille $\Omega = \bigcup_{i\in\mathbb{N}} A_i$ ja $\mu(A_i) < \infty$.

Huomautus. Huomaa, että ehdot ovat samat kuin mitan määritelmässä. Kuitenkin termi *mitta* on varattu kuvauksille, jotka ovat määritelty σ-algebrassa.

Lause 2.4. *[Carathéodoryn laajennuslause] Olkoon $\mathcal{A}_0$ avaruuden Ω algebra ja olkoon $\mathcal{A} = \sigma(\mathcal{A}_0)$.*

Jos kuvaus $\lambda_0\colon \mathcal{A}_0 \to [0,\infty]$ on σ-additiivinen esimitta, on olemassa yksikäsitteinen σ-additiivinen mittalaajennus $\lambda\colon \mathcal{A} \to [0,\infty]$, jolla $\lambda(A) = \lambda_0(A)$ kaikilla $A \in \mathcal{A}_0$.

Todistus. Carathéodoryn laajennuslause todistetaan käyttäen *ulkomitan* käsitettä. Tutkielman laajuuden huomioonottaen emme käsittele todistusta, vaan annamme viitteen todistukseen, joka on käsitelty esimerkiksi kirjassa *Probability with Martingales* [1, s. 195-199]. □

2.3 Mitalliset kuvaukset

Tässä alaluvussa käsittelemme mitallisten kuvausten ominaisuuksia todennäköisyysteorian lähtökohdista. Yleistä mittateoriaa emme käsittele, sillä laajuudessaan se vaatisi syvempää topologian ja funktionaalianalyysin tuntemusta. Esitämme tässä luvussa monotonisen konvergenssin lauseen, joka on erityisen tärkeä integrointiteoriaan liittyvien lauseiden kannalta.

Määritelmä 2.11. Olkoot $(\Omega, \mathcal{A})$ ja $(E, \mathcal{E})$ mitallisia avaruuksia. Tällöin kuvaus $f\colon \Omega \to E$ on mitallinen, jos jokaisella $A \in \mathcal{E}$ pätee $f^{-1}(A) \in \mathcal{A}$.

Tyypillisesti $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, jossa $\mathcal{B}(\mathbb{R}^d)$ on avoimien joukkojen virittämä σ-algebra. Sanotaan, että f on *Borel-mitallinen*, jos $\mathcal{E}$ on topologisen avaruuden Borel-joukkojen kokoelma.

Huomautus. Mitallisten kuvausten määritelmä on analoginen jatkuvien kuvausten topologisen määritelmän kanssa. Olkoot $(S_1, \mathcal{S}_1)$ ja $(S_S, \mathcal{S}_2)$ topologiset avaruudet, joissa $\mathcal{S}_1$ ja $\mathcal{S}_2$ ovat avointen joukkojen kokoelmat. Sanotaan, että kuvaus $f\colon (S_1, \mathcal{S}_1) \to (S_2, \mathcal{S}_2)$ on jatkuva, jos ja vain jos kaikille avoimille $U \in \mathcal{S}_2$ pätee

$$f^{-1}(U) = \{x \in S_1 \colon f(x) \in U\} \in \mathcal{S}_1,$$

eli joukko U on avoin S_1:n topologiassa.

Lause 2.5. *Olkoot $(E, \mathcal{E}), (F, \mathcal{F})$ ja $(G, \mathcal{G})$ mitallisia avaruuksia. Tällöin jos $f\colon E \to F$ ja $g\colon F \to G$ ovat mitallisia kuvauksia, niin $g \circ f\colon E \to G$ on mitallinen kuvaus.*

Todistus. Jos $B \in \mathcal{G}$, niin silloin

$$(g \circ f)^{-1}(B) = f^{-1}(g^{-1}(B)),$$

jolloin kuvauksen Y mitallisuuden perusteella $g^{-1}(B) \in \mathcal{F}$, ja myös kuvauksen f mitallisuuden perusteella $f^{-1}(g^{-1}(B)) \in \mathcal{E}$. □

Lause 2.6. *Olkoon $(\Omega, \mathcal{A})$ mitallinen avaruus. Kuvaus $f\colon \Omega \to \mathbb{R}$ on mitallinen, jos ja vain jos yksi seuraavista ehdoista toteutuu*

(i) $\{x \in \Omega : f(x) < b\} \in \mathcal{A}$,

(ii) $\{x \in \Omega : f(x) \leq b\} \in \mathcal{A}$,

(iii) $\{x \in \Omega : f(x) > b\} \in \mathcal{A}$ *tai*

(iv) $\{x \in \Omega : f(x) \geq b\} \in \mathcal{A}$

kaikilla $b \in \mathbb{R}$. Merkitsemme $\{f < b\} := \{x \in \Omega\colon f(x) < b\}$.

Todistus. Vrt. [1, s. 30] Esimerkiksi $\{x \in \Omega\colon f(x) \leq b\} = f^{-1}((-\infty, b])$. Esimerkin 2.2 perusteella joukko $\{(-\infty, b]\colon b \in \mathbb{R}\}$ virittää Borelin σ-algebran $\mathcal{B}(\mathbb{R})$. Nyt määritelmän 2.11 perusteella kuvaus f on mitallinen. □

Lause 2.7. *Olkoon $(\Omega, \mathcal{A})$ mitallinen avaruus ja $f, g\colon \Omega \to \mathbb{R}$ mitallisia kuvauksia.*

(i) Jos $\lambda \in \mathbb{R}$ ja f on mitallinen, niin λf on mitallinen kuvaus.

(ii) Jos g, h ovat mitallisia kuvauksia, niin $g + h$ on mitallinen kuvaus.

(iii) Jos g, h ovat mitallisia, niin gh on mitallinen kuvaus.

(iv) Jos g, h ovat mitallisia kuvauksia, niin g/h on mitallinen kuvaus.

Todistus. Vrt. [1, s. 30]

(i) Jos $\lambda > 0$, niin

$$\{\lambda f < c\} = \{f < c/\lambda\},$$

joten λf on mitallinen. Vastaavasti tapaukset $\lambda < 0$ ja $\lambda = 0$.

(ii) Olkoon $c \in \mathbb{R}$. On selvää että $g + h > c$ jos ja vain jos on olemassa $q \in \mathbb{Q}$, jolle

$$g > q > c - h.$$

Toisin sanoen nyt

$$\{g + h > c\} = \bigcup_{q \in \mathbb{Q}} (\{g > q\} \cap \{h > c - q\}).$$

(iii) Olkoon $b \in \mathbb{R}$ ja $b \geq 0$. Kuvaus g^2 on mitallinen, sillä

$$\{g^2 < b\} = \{-\sqrt{b} < g < \sqrt{b}\}.$$

Kuvauksien g^2 ja $g + h$ mitallisuudesta seuraa, että

$$gh = \frac{1}{2}((g + h)^2 - g^2 - h^2)$$

on mitallinen.

(iv) Olkoon $g \neq 0$. Nyt

$$\{1/g\} = \begin{cases} \{1/b < g < 0\}, & \text{kun } b < 0, \\ \{-\infty < g < 0\}, & \text{kun } b = 0, \\ \{-\infty < g < 0\} \cup \{1/b < g < \infty\}, & \text{kun } b > 0. \end{cases}$$

Joten $1/g$ on mitallinen, ja näin ollen myös h/g on mitallinen.

□

Lause 2.8. *Jos* $f_n \colon X \to \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ *on mitallinen kaikilla* $n \in \mathbb{N}$ *ja*

$$g = \sup f_n, \quad h = \lim_{n \to \infty} \sup f_n,$$

niin g *ja* h *ovat mitallisia.*

Todistus. Ks. [5, s. 14]. □

Korollaari 2.1. Vrt.[5, s. 15] Olkoon f ja g mitallisia kuvauksia ja $\Omega \to \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$. Tällöin $\max\{f, g\}$ ja $\min\{f, g\}$ ovat mitallisia funktioita. Erityisesti mitallisia ovat funktiot $\max\{f, 0\}$ ja $-\min\{f, 0\}$. Merkitään

$$f^+ := \max\{f, 0\} \quad \text{ja} \quad f^- := -\min\{f, 0\}.$$

Huomautus. Ylläolevaa funktiota f^+ kutsutaan funktion f *positiiviosaksi* ja funktiota f^- funktion f *negatiiviosaksi*. Lisäksi $|f| = f^+ + f^-$ ja $f = f^+ - f^-$.

Määritelmä 2.12. Olkoon Ω joukko ja $A \subset \Omega$. *Indikaattorifunktio* on kuvaus $\mathbb{I}_A \colon \Omega \to \{0, 1\}$, joka on määritelty seuraavasti

$$\mathbb{I}_A(x) := \begin{cases} 1, & \text{jos } x \in A, \\ 0, & \text{jos } x \notin A. \end{cases}$$

Määritelmä 2.13. Funktio $s \colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ on *yksinkertainen*, jos

$$s = \sum_{k=1}^{n} a_k \mathbb{I}_{A_k},$$

missä $A_k = \{x \,|\, s(x) = a_n\}$ ja $A_i \cap A_j = \emptyset$ kaikilla $i, j \in \mathbb{N}$. Yksinkertaisten funktioiden luokkaa merkitsemme $s \in \mathcal{Y}$. Jos funktio on ei-negatiivinen eli $s \geq 0$, niin merkitsemme $s \in \mathcal{Y}^+$.

Määritelmä 2.14 (Yksinkertaisten ei-negatiivisten funktioiden integraali)**.** Olkoon $(\Omega, \mathcal{A}, \mu)$ mitta-avaruus, $A \in \mathcal{A}$ ja s yksinkertainen funktio. Määritellään

$$\mu(s) := \sum_{i=1}^{n} a_i \mu(A_i \cap A).$$

Jos kuvaus $f \colon \Omega \to [0, \infty]$ on mitallinen, niin määritellään

$$\mu(f) := \sup\{\mu(s) \colon s \in \mathcal{Y}^+, 0 \leq s \leq f\} \leq \infty.$$

Tätä integraalia kutsutaan funktion f mittaintegraaliksi mitan μ suhteen. Voidaan ottaa käyttöön merkintä

$$\mu(f) := \int_A f(x)\mu(dx) := \int_A f d\mu.$$

Lause 2.9. *Olkoot funktiot f, g ja joukot A, B mitallisia.*

(i) Jos $0 \leq f \leq g$, niin $\int_A f d\mu \leq \int_A g d\mu$.

(ii) Jos $A \subseteq B$ ja $f \geq 0$, niin $\int_A f d\mu \leq \int_B f d\mu$.

(iii) Jos $f \geq 0$ ja c vakio siten, että $0 \leq c < \infty$, niin

$$\int_A cf d\mu = c \int_A f d\mu.$$

Todistus. [5, s. 19-20]. Väitteet ovat suoraa seurausta määritelmästä 2.14. □

Lause 2.10. *Olkoon* $(\Omega, \mathcal{A})$ *mitta-avaruus ja* $s, t \in \mathcal{Y}^+$. *Silloin*

$$\int_\Omega (s+t)d\mu = \int_\Omega s d\mu + \int_\Omega t d\mu.$$

Todistus. Todistus sivuutetaan ja annetaan viite kirjaan *Real and Complex Analysis* [5, s. 20]. □

Lause 2.11 (Monotonisen konvergenssin lause)**.** *Olkoon* $(\Omega, \mathcal{A}, \mu)$ *mitta-avaruus. Tällöin jos* $(f_n)_{n\in\mathbb{N}}$ *on sellainen jono mitallisia funktioita* $\Omega \to [0, \infty]$, *että jokaisella* $x \in \Omega$

1. $0 \le f_1(x) \le f_2(x) \le \cdots \le \infty$ *ja*
2. $f_n(x) \to f(x)$, *kun* $n \to \infty$,

niin f *on mitallinen, ja*

$$\int_\Omega f_n(x)\mu(dx) \uparrow \int_\Omega f(x)\mu(dx).$$

Todistus. Vrt.[5, s. 21-22] Koska $\int_\Omega f_n \le \int_\Omega f_{n+1}$ kaikilla $n \in \mathbb{N}$, niin on olemassa $a \in [0, \infty]$ siten, että

$$\int_\Omega f_n d\mu \to a, \quad \text{kun } n \to \infty.$$

Lauseen 2.8 perusteella kuvaus f on mitallinen. Koska $f_n \le f$, niin jokaisella $n \in \mathbb{N}$ pätee $\int f_n \le \int f$. Siis

$$a \le \int_\Omega f d\mu.$$

Olkoon s yksinkertainen funktio siten, että $0 \le s \le f$, ja olkoon $c \in (0, 1)$. Määritellään jono

$$E_n = \{x \colon f_n(x) \ge cs(x)\}, \quad \text{kun } n \in \mathbb{N}.$$

Nyt jokainen E_n on mitallinen ja $E_n \uparrow \Omega$. Nähdäksemme tämän yhtäsuuruuden, olkoon $x \in \Omega$. Jos $f(x) = 0$, niin $x \in E_1$, ja jos taas $f(x) > 0$, niin $cs(x) < f(x)$, kun $c < 1$. Siis $x \in E_n$ jollakin n. Myös

$$\int_\Omega f_n d\mu \ge \int_{E_n} f_n d\mu \ge c \int_{E_n} d\mu, \quad n \in \mathbb{N}.$$

Kun $n \to \infty$, niin lauseiden 2.10 ja 2.2 nojalla pätee

$$a \ge c \int_\Omega s d\mu$$

kaikilla $c < 1$. Siis

$$a \geq \int_\Omega s d\mu$$

jokaisella yksinkertaisella funktiolla $0 \leq s \leq f$, jolloin

$$a \geq \int_\Omega f d\mu.$$

□

Määritelmä 2.15 (Integroituva funktio)**.** Olkoon $(\Omega, \mathcal{A}, \mu)$ mitallinen avaruus. Kun $f \in \mathcal{A}$, niin sanomme että f on μ-integroituva ja merkitsemme $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$, jos

$$\mu(|f|) = \mu(f^+) + \mu(f^-) < \infty.$$

Voimme määritellä, että

$$\int f d\mu := \mu(f) := \mu(f^+) - \mu(f^-).$$

Huomautus. Huomioidaan, että kun $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$, niin $|\mu(f)| \leq \mu(|f|)$.

Lemma 2.2. Olkoon $(\Omega, \mathcal{A}, \mu)$ mitta-avaruus. Lisäksi olkoon $f\colon X \to [0, \infty]$ mitallinen funktio, ja

$$F(A) = \int_A f(x)\mu(dx),$$

kun $A \in \mathcal{A}$. Tällöin $F\colon \mathcal{A} \to [0, \infty]$ on mitta, ja

$$\int_\Omega g(x)F(dx) = \int_\Omega (gf)(x)\mu(dx)$$

jokaisella mitallisella $g\colon X \to [0, \infty]$.

Todistus. Ks. [5, s. 23-24] Lause seuraa suoraan monotonisen konvergenssin lauseesta 2.11. □

2.4 Lebesgue-Stieltjesin -mitta

Ennen kuin määrittelemme Lebesgue-Stieltjesin mitan, käymme läpi muutamia peruskäsitteitä ja määritelmiä. Kun $a, b \in \mathbb{R}^n$ ja $a = (a_1, \ldots, a_n)$ sekä $b = (b_1, \ldots, b_n)$, niin määritellään

$$(a, b] = \{x \in \mathbb{R}^n \mid \forall i \in \{1, \ldots, n\}\colon a_i < x_i \leq b_i\}, \tag{2.2}$$

$$(-\infty, a) = \{x \in \mathbb{R}^n \mid \forall i \in \{1, \ldots, n\}\colon a_i > x_i\} \tag{2.3}$$

ja

$$(-\infty, a] = \{x \in \mathbb{R}^n \mid \forall i \in \{1, \ldots, n\}\colon a_i \geq x_i\}. \tag{2.4}$$

Lisäksi $a \leq b$ tarkoittaa, että $a_i \leq b_i$ kaikilla $i \in \{1, \ldots, n\}$. Kuitenkin tässä tutkielmassa pitäydytään tapauksessa $n = 1$.

Määritelmä 2.16. Avaruuden $\mathbb{R}$ Borelin mittaa μ kutsutaan *Lebesgue-Stieltjesin mitaksi*, jos $\mu(I) < \infty$, kun I on rajoitettu väli ja on muotoa $(a, b]$.

Määritelmä 2.17. Olkoon $F\colon \mathbb{R} \to \mathbb{R}$ funktio.

(i) Funktio F on *kasvava*, kun $a, b \in \mathbb{R}$ ja $a < b$, niin $F(a) \leq F(b)$.

(ii) Funktio F on *oikealta jatkuva*, kun $\lim_{x \to x_0+} F(x) = F(x_0)$.

(iii) Funktio F on *kertymäfunktio*, jos se on kasvava ja oikealta jatkuva.

Lause 2.12. *Olkoon μ Lebesgue-Stieltjesin mitta avaruudessa $\mathbb{R}$. Lisäksi olkoon kuvaus $F\colon \mathbb{R} \to \mathbb{R}$ määritelty seuraavasti*

$$F(b) - F(a) = \mu((a, b]).$$

Tällöin F on kertymäfunktio.

Todistus. Vrt. [4, s. 23] Olkoon $a, b \in \mathbb{R}$. Jos $a < b$, niin kaikilla $F(b) - F(a) = \mu((a, b]) \geq 0$. Jos $(x_n)_{n \in \mathbb{N}}$ on jono, jolle $x_1 > x_2 > \cdots \to x$, niin lauseen 2.2 perusteella $F(x_n) - F(x) = \mu((x, x_n]) \to 0$. □

Huomautus. Caratheodoryn laajennuslauseessa mitallisuutta käsitellään kompaktissa reaaliavaruudessa $\bar{\mathbb{R}}$. Laajennamme siis kertymäfunktion lähtö- ja maalijoukkoa määrittelemällä

(i) $F(\infty) = \lim_{x \to \infty} F(x)$ ja

(ii) $F(-\infty) = \lim_{x \to -\infty} F(x)$

joissa raja-arvot ovat olemassa funktion monotonisuuden takia. Lisäksi määrittelemme

(i) $\mu((a, b]) = F(b) - F(a)$ kaikilla $a, b \in \bar{\mathbb{R}}$ kun $a < b$ ja

(ii) $\mu([-\infty, b]) = F(b) - F(-\infty) = \mu((-\infty, b])$,

jolloin μ on määritelty kaikilla vasemmalta puoliavoimilla väleillä joukossa $\bar{\mathbb{R}}$. Lisäksi jos $I_1, \ldots, I_k$ on erillisiä vasemmalta puoliavoimia välejä joukossa $\bar{\mathbb{R}}$, niin määrittelemme $\mu\left(\bigcup_{j=1}^k I_j\right) = \sum_{j=1}^k \mu(I_j)$. Nyt μ on äärellisesti additiivinen joukossa $\bar{\mathbb{R}}$.

Jotta Caratheodoryn laajennuslauseen 2.4 ehdot ovat voimassa, niin todistamme vielä, että esimitta μ on numeroituvasti additiivinen algebrassa $\mathcal{A}_0(\bar{\mathbb{R}})$. Tällöin esimitalla μ on yksikäsitteinen laajennus avaruuteen $\sigma(\mathcal{A}_0(\bar{\mathbb{R}})) = \mathcal{B}(\mathbb{R})$.

Lemma 2.3. Esimitta μ on numeroituvasti additiivinen algebrassa $\mathcal{A}_0(\bar{\mathbb{R}})$.

Todistus. Vrt. [4, s. 23-24] Oletetaan, että μ on äärellinen ja jono $\{A_n : n \in \mathbb{N}\} \subseteq \mathcal{A}_0(\bar{\mathbb{R}})$ puoliavoimia välejä, joilla $A_n \downarrow \emptyset$. Jos $(a, b]$ on yksi jonon $(A_n)_{n\in\mathbb{N}}$ väleistä, niin suoraan F jatkuvuuden perusteella

$$\mu((a', b]) = F(b) - F(a') \to F(b) - F(a) = \mu((a, b]), \text{ kun } a' \to a.$$

Joten voimme etsiä joukot $B_n \in \mathcal{A}_0(\bar{\mathbb{R}})$, joiden sulkeumat $\bar{B_n}$ kuuluvat joukkoihin A_n, joissa mitta $\mu(B_n)$ mukailee/arvioi mittaa $\mu(A_n)$. Jos $\epsilon > 0$ on annettu, niin mitan μ äärellisyyden seurauksena voimme valita B_n siten, että $\mu(A_n) - \mu(B_n) < \frac{\epsilon}{2^n}$. Nyt $\bigcap_{n=1}^{\infty} \bar{B}_n = \emptyset$, joka seuraa siitä, että $\bigcap_{k=1}^{n} \bar{B}_k = \emptyset$, kun n on tarpeeksi suuri. Tämä nähdään helposti, kun joukkojen erotus $\bar{\mathbb{R}} - \bar{B}_n$ on kompaktin joukon $\bar{\mathbb{R}}$ peitteenä, jolloin on olemassa äärellinen alipeite, jolla $\bigcup_{k=1}^{n}(\bar{\mathbb{R}} - \bar{B}_k) = \bar{\mathbb{R}}$, jollain $n \in \mathbb{N}$. Nyt

$$\begin{aligned}
\mu(A_n) &= \mu\left(A_n - \bigcap_{k=1}^{n} B_k\right) + \mu\left(\bigcap_{k=1}^{n} B_k\right) \\
&= \mu\left(A_n - \bigcap_{k=1}^{n} B_k\right) \\
&\leq \mu\left(\bigcup_{k=1}^{n} (A_k - B_k)\right), \quad \text{koska } A_n \subset A_{n-1} \subset \cdots \subset A_1 \\
&\leq \sum_{k=1}^{n} \mu(A_k - B_k) \\
&< \epsilon.
\end{aligned}$$

Joten $\mu(A_n) \to 0$.

Nyt jos $F(\infty) - F(-\infty) = \infty$, niin määritellään

$$F_n(x) = \begin{cases} F(x), & \text{kun } |x| \leq n, \\ F(n), & \text{kun } x \geq n, \\ F(-n), & \text{kun } x \leq -n. \end{cases}$$

Jos μ_n on mitallinen funktio, joka vastaa funktiota F_n, silloin $\mu_n \leq \mu$ ja $\mu_n \to \mu$. Olkoon $A_n \uparrow A$, silloin $\mu(A) \geq \sum_{n=1}^{\infty} \mu(A_n)$. Jos $\sum_{n=1}^{\infty} \mu(A_n) = \infty$, niin ei ole todistettavaa. Jos $\sum_{n=1}^{\infty} \mu(A_n) < \infty$, niin

$$\begin{aligned}
\mu(A) &= \lim_{n\to\infty} \mu_n(A) \\
&= \lim_{n\to\infty} \sum_{k=1}^{\infty} \mu_n(A_k),
\end{aligned}$$

koska μ_n on äärellinen. Nyt koska $\sum_{k=1}^{\infty} \mu(A_k) < \infty$, niin

$$\begin{aligned} 0 &\leq \mu(A) - \sum_{k=1}^{\infty} \mu(A_k) \\ &= \lim_{n\to\infty} \sum_{k=1}^{\infty} (\mu_n(A_k) - \mu(A_k)) \\ &\leq 0, \qquad \text{koska } \mu_n \leq \mu. \end{aligned}$$

Siis esimitta μ on numeroituvasti äärellinen. □

Lause 2.13. *Olkoon $F\colon \mathbb{R} \to \mathbb{R}$ kertymäfunktio, ja olkoon mitallinen kuvaus $\mu((a,b]) = F(b) - F(a)$, kun $a < b$ kaikilla $a, b \in \mathbb{R}$. Tällöin funktiolla μ on olemassa yksikäsitteinen laajennus, joka on Lebesguen-Stieltjesin mitta avaruudessa $\mathbb{R}$.*

Todistus. [4, s. 24-25] Lemman 2.3 nojalla μ on numeroituvasti additiivinen. Olkoon $\mathcal{A}_0(\mathbb{R})$ kaikkien puoliavointen välien algebra ja μ kuten lauseen 2.12 huomautuksessa. Nyt puoliavoin väli $(a, b]$ on määritelty algebrassa $\mathcal{A}_0(\bar{\mathbb{R}})$ siten, että $a, b \in \mathbb{R}$ tai $a = -\infty$ ja $b \in \mathbb{R}$. Puoliavoin väli $(a, \infty]$ taas on määritelty siten, että $a \in \mathbb{R}$ tai $a = -\infty$. Lisäksi μ on σ-äärellinen $\mathcal{A}_0(\mathbb{R})$:ssä, sillä $\mu((-n, n]) < \infty$. Caratheodoryn lauseen 2.4 nojalla, funktiolla μ on yksikäsitteinen laajennus joukolle $\mathcal{B}(\mathbb{R})$. Tämä laajennus on Lebesguen-Stieltjesin mitta, sillä $\mu((a, b]) = F(b) - F(a)$, kun $a < b$ kaikilla $a, b \in \mathbb{R}$. □

Huomautus. Nyt voimme konstruoida Lebesguen mitan. Olkoon μ äärellinen, jolloin kertymäfunktio F on rajoitettu. Koska kertymäfunktiota F voidaan muokata integroimisvakiolla, niin voidaan asettaa $F(-\infty) = 0$. Olkoon $f\colon \mathbb{R} \to \mathbb{R}$ ja $f \geq 0$. Olkoon lisäksi funktio f Riemann-integroituva millä tahansa äärellisellä välillä, niin määrittelemme

$$F(x) - F(0) = \int_0^x f(t)dt, \quad \text{kun } x > 0,$$

$$F(0) - F(x) = \int_x^0 f(t)dt \quad \text{kun } x < 0,$$

niin silloin F on kertymäfunktio. Erityisesti nyt

$$\mu((a, b]) = \int_a^b f(x)dx.$$

Olkoon $f(x) = 1$ kaikilla $x \in \mathbb{R}$ ja $F(x) = x$. Tällöin $\mu((a, b]) = b - a$. Kutsumme tätä funktiota *Lebesguen mitaksi* Borelin joukossa $\mathcal{B}(\mathbb{R})$.

3 Todennäköisyysteoriaa

Määrittelimme todennäköisyysmitan ja todennäköisyysavaruuden käsitteen jo luvussa 2 määritelmässä 2.5. Tässä luvussa käsittelemme todennäköisyysteoriaa, joka voidaan mieltää mittateorian erityistapauksena. Todennäköisyysteoria antaa työkalut todennäköisyyslaskentaan, joka antaa edellytykset erilaisiin matemaattisiin sovellutuksiin ja tilastolliselle päättelylle. Todennäköisyysteorian osuus perustuu kirjoihin *Probability with Martingales* [1], *Probability and Measure*[10] ja *Real Analysis and Probability* [4].

3.1 Satunnaismuuttuja ja todennäköisyysjakauma

Määritelmä 3.1. Olkoon $(\Omega, \mathcal{A}, P)$ todennäköisyysavaruus. *Satunnaismuuttuja* X on kuvaus mitalliseen avaruuteen $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, jossa $X\colon \Omega \to \mathbb{R}$. Jos X on kuvaus mitalliseen avaruuteen $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, niin kutsumme satunnaismuuttujaa X *satunnaisvektoriksi*.

Määritelmä 3.2. Olkoot $X_1, X_2, \ldots, X_n$ satunnaismuuttujia todennäköisyysavaruudessa $(\Omega, \mathcal{A}, P)$. Satunnaismuuttujat $X_1, X_2, \ldots, X_n$ ovat *riippumattomia*, jos ja vain jos jokaisella joukolla $B_1, B_2, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ pätee

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

Määritelmä 3.3. Olkoon $X\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satunnaismuuttuja, ja olkoon P todennäköisyysmitta avaruudessa $(\Omega, \mathcal{A})$. Silloin todennäköisyysmittaa

$$P_X(B) := (P \circ X^{-1})(B), \quad \text{kaikilla } B \in \mathcal{S}$$

kutsutaan satunnaismuuttujan X todennäköisyysjakaumaksi.

Määritelmä 3.4. Olkoon X satunnaismuuttuja. Satunnaismuuttuja X on *absoluuttisesti jatkuva*, jos ja vain jos on olemassa ei-negatiivinen Borel-mitallinen funktio f, siten että

$$F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in \mathbb{R}.$$

Kutsumme funktiota f satunnaismuuttujan X *tiheysfunktioksi*. Koska $F(x) \to 1$, kun $x \to \infty$, niin $\int_{-\infty}^{\infty} f(x)dx = 1$.

Lause 3.1. *Olkoon X absoluuttisesti jatkuva satunnaismuuttuja ja f sen tiheysfunktio. Tiheysfunktiolle pätee*

$$P_X(B) = \int_B f(x)dx, \quad \textit{jokaisella } B \in \mathcal{B}(\mathbb{R}).$$

Todistus. Vrt. [4, s. 210] Määritelmän 2.14 perusteella $\mu(B) = \int_B f(x)dx$, kun $B \in \mathcal{B}(\mathbb{R})$, jolloin pätee $\mu((a,b]) = F(b) - F(a)$, kun $a < b$. Siis μ on Lebesgue-Stieltjes -mitta funktiolle F. Siis $\mu = P_X$. □

Lause 3.2. *Olkoon $f\colon \mathbb{R}^d \to \mathbb{R}$ ei-negatiivinen Borel-mitallinen funktio ja $\int_{\mathbb{R}^d} f(x)dx = 1$. Tällöin funktio f on todennäköisyyden $P\colon \mathbb{R}^d \to [0,1]$ tiheysfunktio.*

Todistus. Olkoon $P(B) := \int_A f(x)dx$, kun $B \in \mathcal{B}(\mathbb{R}^d)$. Nyt lemman 2.2 perusteella P on mitta. Oletuksen perusteella $\int_{\mathbb{R}^d} f(x)dx = 1$, jolloin P on todennäköisyysmitta. □

Esimerkki 3.1. Tyypillisiä tiheysfunktioita ovat esimerkiksi:

1. Tasajakauma välillä $[a,b]$:

$$f(x) = \begin{cases} (b-a)^{-1}, & \text{kun } a \le x \le b, \\ 0, & \text{muulloin.} \end{cases}$$

2. Normaalijakauma:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \text{kun } \sigma > 0 \text{ ja } m \in \mathbb{R}.$$

3.2 Odotusarvo

Määritelmä 3.5. Olkoon $(\Omega, \mathcal{A}, P)$ todennäköisyysavaruus. Satunnaismuuttujan X *odotusarvo* $\mathbb{E}(X)$ on silloin

$$\mathbb{E}(X) := \int_\Omega X(x)P(dx).$$

Jos $\mathbb{E}(|X|) < \infty$, niin X on integroituva ja määrittelemme

$$\mathbb{E}(X) := \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Lause 3.3. *Olkoon X satunnaismuuttuja avaruudesssa $(\Omega, \mathcal{A}, P)$ kertymäfunktiolla F. Olkoon $g\colon \mathbb{R} \to \mathbb{R}$ Borel-mitallinen funktio. Tällöin*

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)F(dx) := \int_{\mathbb{R}} g(x)P_X(dx),$$

jos jompi kumpi yhtälön puolista on olemassa.

Todistus. Vrt.[4, s. 224] Olkoon g indikaattorifunktio $\mathbb{I}_B$, jossa $B \in \mathcal{B}(\mathbb{R})$. Silloin

$$\mathbb{E}(g(X)) = \mathbb{E}(\mathbb{I}_B \circ X) = \mathbb{E}(\mathbb{I}_{\{X \in B\}}) = P_X(B) = \int_{\mathbb{R}} g(x) P_X(dx),$$

jolloin $\mathbb{E}(g(X))$ ja $\int_{\mathbb{R}} g(x) P_X(dx)$ ovat olemassa ja ovat yhtä suuria.

Olkoon nyt g ei-negatiivinen yksinkertainen funktio. Lisäksi olkoon $g(x) = \sum_{j=1}^{n} x_j \mathbb{I}_B(x)$ ja joukot B_j erillisiä Borelin joukossa $\mathcal{B}(\mathbb{R})$. Silloin

$$\begin{aligned}
\mathbb{E}(g(X)) &= \sum_{j=1}^{n} x_j \mathbb{E}(\mathbb{I}_{B_j} \circ X) \\
&= \sum_{j=1}^{n} x_j \int_{\mathbb{R}} \mathbb{I}_{B_j} P_X(dx) \\
&= \int_{\mathbb{R}} \left(\sum_{j=1}^{n} x_j \mathbb{I}_{B_j} \right) P_X(dx), \quad \text{koska } g \geq 0 \\
&= \int_{\mathbb{R}} g(x) P_X(dx).
\end{aligned}$$

Jos g on ei-negatiivien Borel-mitallinen funktio, olkoon $g_1, g_2, \ldots, g_n$ ei-negatiivisia yksinkertaisia funktioita, joilla $g_n \uparrow g$. Koska todistimme äsken, että

$$\mathbb{E}(g_n(X)) = \int_{\mathbb{R}} g_n P_X(dx),$$

niin monotonisen konvergenssin lauseen 2.11 perusteella

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g P_X(dx).$$

Jos $g = g^+ - g^-$ on mielivaltainen Borel-mitallinen funktio, niin

$$\begin{aligned}
\mathbb{E}(g(X)) &= \mathbb{E}(g^+(X)) - \mathbb{E}(g^-(X)) \\
&= \int_{\mathbb{R}} g^+(x) P_X(dx) - \int_{\mathbb{R}} g^-(x) P_X(dx) \\
&= \int_{\mathbb{R}} g(x) P_X(dx).
\end{aligned}$$

Jos $\mathbb{E}(g(X))$ on olemassa, niin $\mathbb{E}(g^-(X))$ on äärellinen, jolloin myös $\int_{\mathbb{R}} g^-(x) P_X(dx)$ on äärellinen, ja siis integraali $\int_{\mathbb{R}} g(x) P_X(dx)$ on olemassa. Sama pätee toisinpäin, jos integraali $\int_{\mathbb{R}} g(x) P_X(dx)$ on olemassa, niin siitä seuraa myös odotusarvon $\mathbb{E}(g(X))$ olemassaolo. □

Lause 3.4. *Olkoon X on satunnaismuuttuja, jolla on tiheysfunktio f. Tällöin*

$$\int g(x) F(dx) = \int g(x) f(x) dx.$$

Todistus. Sivuutetaan, ks. [4, s. 225], vastaavasti kuin edellinen todistus. □

Seuraavaksi käsittelemme tutkielman luonteesta poiketen todennäköisyysteoriaa myös todennäköisyysavaruudessa $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P)$. Tarkempi käsittely vaatisi tuloavaruuden ja tulomitan käsitteen määrittelyn sekä Fubinin-Tonellin lauseen todistuksen. Lisäksi olemme todistaneet Lebesgue-Stieltjesin mitan vain mitalliselle avaruudelle $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Seuraavat lause motivoi alaluvussa 3.3 esiteltävän yhdistetyn jakauman momenttien määrittelyn.

Lause 3.5. *Olkoon $X_1, X_2, \ldots, X_n$ riippumattomia satunnaismuuttujia todennäköisyysavaruudessa $(\Omega, \mathcal{A}, P)$. Jos X_i on ei-negatiivinen satunnaismuuttuja tai odotusarvo $\mathbb{E}(X_i)$ on äärellinen, niin $\mathbb{E}(X_1 X_2 \cdots X_n)$ on olemassa ja $\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_n)$.*

Todistus. Ks. [4, s. 227-228]. □

Määritelmä 3.6. Olkoon $X, Y \colon \Omega \to \mathbb{R}$ satunnaismuuttujia, joille $\mathbb{E}(X), \mathbb{E}(Y), \mathbb{E}(XY) < \infty$. Tällöin

$$Cov(X,Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{3.1}$$

on satunnaismuuttujien X ja Y *kovarianssi*, ja

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = Cov(X,X) \tag{3.2}$$

on satunnaismuuttujan X varianssi.

Määritelmä 3.7. Olkoon $X \colon \Omega \to \mathbb{R}^d$ ja $Y \colon \Omega \to \mathbb{R}^k$ satunnaisvektoreita. Tällöin

$$\mathbb{E}(X) := (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d)) \tag{3.3}$$

on satunnaisvektorin X *odotusarvovektori* ja

$$Cov(X,Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^T] \tag{3.4}$$

on satunnaisvektorien X ja Y *kovarianssimatriisi*.

3.3 Momentit ja kumulantit generoiva funktio

Määritelmä 3.8. Olkoon X satunnaismuuttuja avaruudessa $(\Omega, \mathcal{A}, P)$. Jos $k \in \mathbb{N}$, niin

(i) satunnaismuuttujan X:n k:nnes *momentti* on $\mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k F(dx)$, ja k:nnes *absoluuttinen momentti* on $\mathbb{E}(|X|^k) = \int_{-\infty}^{\infty} |x|^k F(dx)$.

(ii) satunnaismuuttujan X:n k:nnes *keskusmomentti* on $\mathbb{E}((X - \mathbb{E}(X))^k)$, ja k:nnes *absoluuttinen keskusmomentti* on $\mathbb{E}(|X - \mathbb{E}(X)|^k)$.

Keskusmomentit ovat olemassa ainoastaan silloin, kun odotusarvo $\mathbb{E}(X)$ on äärellinen.

Lemma 3.1. Jos $k > 0$ ja $\mathbb{E}(X^k)$ on äärellinen, niin silloin myös $\mathbb{E}(X^j)$ on äärellinen jokaisella $0 < j < k$.

Todistus. Ks. [4, s. 226] □

Lause 3.6. *Jos $n > 1$ on positiivinen kokonaisluku, odotusarvo $\mathbb{E}(X^{k-1})$ on äärellinen ja odotusarvo $\mathbb{E}(X^n)$ on olemassa, niin silloin*

$$\mathbb{E}((X - \mathbb{E}(X))^n) = \sum_{k=0}^{n} \binom{n}{k} (-\mathbb{E}(X))^{n-k} \mathbb{E}(X^k).$$

Erityisesti, kun $\mathbb{E}(X)$ on äärellinen, niin

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Todistus. Todistus sivuutetaan ja annetaan viite [4, s. 226]. □

Määritelmä 3.9. Olkoon Y satunnaismuuttuja, jolla on olemassa tiheysfunktio f. Satunnaismuuttujan Y *momentit generoiva funktio* on kuvaus $M_Y \colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, joka määräytyy ehdosta

$$M_Y(s) = \mathbb{E}(e^{sY}), \tag{3.5}$$

jos lausekkeen (3.5) oikeanpuoleinen odotusarvo on olemassa jossain nollan ympäristössä. Toisin sanoen on olemassa $h > 0$ siten, että kaikilla s:n arvoilla $-h < s < h$, odotusarvo $\mathbb{E}(e^{sY})$ on olemassa. Lisäksi satunnaismuuttujan Y *kumulantit generoiva funktio* on kuvaus $C_Y \colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, joka määräytyy ehdosta

$$C_Y(s) = \log\left(\mathbb{E}(e^{sY})\right). \tag{3.6}$$

Esimerkki 3.2. Satunnaismuuttujan Y momentit generoiva funktio on

$$M_Y(s) = \sum_{k=1}^{\infty} e^{sy} P\{Y = y\}, \quad \text{kun } Y \text{ on diskreetti, ja}$$

$$M_Y(s) = \int_{-\infty}^{\infty} e^{sy} f(y) dy, \quad \text{kun } Y \text{ on jatkuva.}$$

Huomautus. Satunnaismuuttuja Y on diskreetti, kun se voi saada arvoja vain äärellisesti tai numeroituvan äärettömästi.

Lause 3.7. *Jos satunnaismuuttujan Y momentit generoiva funktio $M(s)$ on olemassa välillä $[-s_0, s_0]$, kun $s_0 > 0$, on sen kaikkien asteiden derivaatat olemassa, kun $s = 0$ ja*

$$M^{(k)}(0) = \mathbb{E}(X^k), \quad \textit{kun } k \in \mathbb{Z}_+.$$

Todistus. Olkoon $M(s) = \mathbb{E}(e^{sY})$ satunnaismuuttujan Y momentit generoiva funktio.

Todistetaan jatkuva tapaus. Oletetaan, että satunnaismuuttuja Y on jatkuva. Tämän seurauksena saadaan

$$\begin{aligned}\frac{d}{ds}M(s) &= \frac{d}{ds}\int_{-\infty}^{\infty} e^{sy}f(y)dy \\ &= \int_{-\infty}^{\infty}(\frac{d}{ds}e^{sy})f(y)dy \\ &= \int_{-\infty}^{\infty}(ye^{sy})f(y)dy \\ &= \mathbb{E}(Ye^{sY}).\end{aligned}$$

Näin ollen

$$M'(0) = \mathbb{E}(Ye^{0Y}) = \mathbb{E}(Y).$$

Kun jatketaan samalla tavalla, päädytään tulokseen

$$M^{(k)}(0) = \frac{d^k}{ds^k}M(0) = \mathbb{E}(Y^k e^{0y}) = \mathbb{E}(Y^k).$$

□

Huomautus. Odotusarvo ja varianssi voidaan määritellä myös kumulantit generoivan funktion kautta. Koska $C(s) = \log M(s)$ ja $M(0) = 1$, niin odotusarvo voidaan määritellä kumulantit generoivan funktion 1. derivaatan avulla seuraavasti

$$C'_Y(0) = \frac{M'(0)}{M(0)} = \mathbb{E}(Y),$$

ja varianssi 2. derivaatan avulla seuraavasti

$$\begin{aligned}C''_Y(0) &= \frac{M''(0)M(0) - M'(0)^2}{M(0)^2} \\ &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= Var(Y).\end{aligned}$$

Seuraavat lauseet ovat käyttökelpoisia varsinkin vakuutusmatemaattisissa sovelluksissa, kun mallinnetaan esimerkiksi kokonaisvahinkomäärää tietylle vahinkoluokalle.

Lause 3.8. *Olkoon $X_1, X_2, \ldots, X_n$ riippumattomia satunnaismuuttujia, joiden momentit generoivat funktiot ovat $M_{X_1}(s), M_{X_2}(s), \ldots, M_{X_n}(s)$. Tällöin*

$$M_{X_1+X_2+\cdots+X_n}(s) = M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s).$$

Todistus. Lauseen 3.5 nojalla

$$\begin{aligned} M_{X_1+X_2+\cdots+X_n}(s) &= \mathbb{E}(e^{s(X_1+X_2+\cdots+X_n)}) \\ &= \mathbb{E}(e^{sX_1}e^{sX_2}\cdots e^{sX_n}) \\ &= \mathbb{E}(e^{sX_1})\mathbb{E}(e^{sX_2})\cdots\mathbb{E}(e^{sX_n}) \\ &= M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s). \end{aligned}$$

□

Lause 3.9. *Olkoon Y ja $Z_1 + Z_2 + \cdots + Z_K$ satunnaismuuttujia sekä $Y = Z_1 + Z_2 + \cdots + Z_K$. Olkoon lisäksi S kertymäfunktio. Muuttujaa Y kutsutaan* yhdistetyksi satunnaismuuttujaksi *parametrillä (K, S), jos*

(i) $K, Z_1, Z_2, \ldots$ ovat riippumattomia

(ii) muuttujien $Z_1, Z_2, \ldots, Z_K$ kertymäfunktio on S.

Olkoon nyt M_K muuttujan K momentit generoiva funktio ja C_Z muuttujan Z kumulantit generoiva funktio. Silloin Y:n momentit generoiva funktio M_Y määräytyy ehdosta

$$M_X(s) = M_K(C_Z(s)),$$

missä $s \in \mathbb{R}$ ja $M_X(s) = \infty$, jos $C_Z(s) = \infty$.

Todistus. Lauseen 3.8 nojalla

$$\begin{aligned} M_Y(s) &= \sum_{k=0}^{\infty} \mathbb{E}(e^{sY}\mathbb{I}(K=k)) = \sum_{k=0}^{\infty} P_K \mathbb{E}(e^{s(Z_1+\cdots+Z_k)}) \\ &= \sum_{k=0}^{\infty} P_k M_Z(s)^k = \sum_{k=0}^{\infty} P_k e^{kC_Z(s)} = M_K(C_Z(s)). \end{aligned}$$

□

4 Tilastollisia esitietoja

Tässä luvussa käymme läpi matemaattisiin sovelluksiin liittyviä käsitteitä. Erityistä huomiota kiinnitämme eksponentiaalisen jakaumaperheen määritelmään ja sen ominaisuuksiin. Eksponentiaalinen jakaumaperhe on keskeisessä roolissa yleistettyjen lineaaristen mallien teoriassa, sillä selitettävän muuttujan jakauman oletetaan kuuluvan johonkin eksponentiaaliseen jakaumaperheeseen. Seuraava luku perustuu kirjoihin *Foundations of Linear and Generalized Linear Models* [3], *The Theory of Exponential Dispersion Models and Analysis of Deviance* [6], *Non-Life Insurance Pricing with Generalized Linear Models* [2], *Correlated Data Analysis: Modeling, Analytics, and Applications* [7] ja julkaisuun *A note on the overdispersed Poisson family* [8].

4.1 Eksponentiaaliset jakaumaperheet

Eksponentiaaliset perheet ovat tärkeä osa yleistettyjä lineaaristen mallien teoriaa. Yleistetyissä lineaarisissa malleissa selitettävän muuttujan jakauman oletetaan kuuluvan johonkin eksponentiaaliseen jakaumaperheeseen.

Määritelmä 4.1. Oletetaan, että

(i) $\Theta \subset \mathbb{R}$ ja $\Phi \subset]0, \infty[$ on epätyhjä joukko,

(ii) $b\colon \Theta \to \mathbb{R}$ on kahdesti derivoituva funktio, jonka ensimmäisellä derivaatalla on käänteisfunktio $b'^{-1}\colon b'(\Theta) \to \Theta$,

(iii) $c\colon \mathbb{R} \times \Phi \to [0, \infty[$ on ensimmäisen muuttujan suhteen mitallinen funktio. Siis kiinnitetään $\phi \in \Phi$, niin $\{y \in \mathbb{R} \,|\, c(y, \phi) \in B\} \in \mathcal{B}(\mathbb{R})$ kaikilla $B \in \mathcal{B}(\mathbb{R})$,

(iv) $\upsilon\colon \mathcal{B}(\mathbb{R}) \to [0, \infty]$ on σ-äärellinen mitta $\upsilon\colon \mathcal{B}(\mathbb{R}) \to [0, \infty]$.

Asetetaan lisäksi ehto, että Θ, Φ, b, c ja υ on valittava siten, että kun kiinnitetään mielivaltaiset parametrit $\theta \in \Theta$ ja $\phi \in \Phi$, niin

$$\int_B f(y, \theta, \phi) d\upsilon(y) = 1,$$

jossa funktio $f\colon \mathbb{R} \to [0, \infty[$ on määritelty asettamalla

$$f(y, \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{\phi}\right) c(y, \phi) \quad \text{kaikilla } y \in \mathbb{R}. \tag{4.1}$$

Tällöin funktio $f(y, \theta, \phi)$ on *eksponentiaalisen jakaumaperheen* tiheysfunktio.

Esimerkki 4.1. Poisson -jakauman pistetodennäköisyysfunktio on muotoa

$$f(y_i, \mu_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}.$$

Voidaan ottaa puolittain logaritmi, jolloin muokkaamme yhtälöä eksponentiaalisen jakaumaperheen vaatimaan muotoon

$$\log f(y_i, \mu_i) = \log\left(e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}\right) = \log\left(e^{-\mu_i} \mu_i^{y_i}\right) - \log(y_i).$$

Nyt siis

$$f(y_i, \mu_i) = \exp\left(y_i \log(\mu_i) - \mu_i - \log(y_i!)\right).$$

Eli saadaan $\theta = \log(\mu_i) \Leftrightarrow \mu_i = e^\theta$, jolloin $b(\theta) = e^\theta$. Lisäksi $\phi = 1$ ja $c(y_i, \phi) = (y_i)^{-1}$, jolloin

$$f(y_i, \theta, \phi) = c(y_i, \phi) \exp\left(y_i\theta - b(\theta)\right).$$

Määritelmä 4.2. Satunnaismuuttujan Y jakauma P_Y kuuluu *eksponentiaaliseen jakaumaperheeseen*, jos on olemassa sellaiset parametrit $\theta \in \Theta$ ja $\phi \in \Phi$, että kaikilla $B \in \mathcal{B}(\mathbb{R})$

$$P_Y(B) = \int_B f(y, \theta, \phi) d\upsilon(y),$$

jossa funktio $f\colon \mathbb{R} \to [0, \infty[$ on määritelty kaavan 4.1 mukaisesti.

Jos mitta υ on Lebesguen mitta, niin kyseessä on jatkuva jakauma, jonka tiheysfunktio on f. Jos mitta υ on lukumäärämitta, niin silloin kyseessä on diskreetti jakauma, jonka pistetodennäköisyysfunktio on f. Muuttujaa θ kutsutaan luonnolliseksi parametriksi ja muuttujaa ϕ kutsutaan dispersioparametriksi. Funktiota b voidaan luonnehtia kumulantit generoivana funktiona.

Taulukko 1: Eksponentiaalisia jakaumaperheitä

	Normaali	Poisson	Gamma
arvojoukko y	$\mathbb{R}$	$\{0, \phi, 2\phi, \dots\}$	$(0, \infty)$
parametrijoukko Θ	$\mathbb{R}$	$\mathbb{R}$	$(-\infty, 0)$
parametrijoukko Φ	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$
$b(\theta)$	$\theta^2/2$	e^θ	$-\log(-\theta)$
$c(y, \phi)$	$2\pi\phi e^{-y^2/2\phi}$	$\phi^{-y/\phi}/(y\phi)!$	$(y/\phi)^{1/\phi}(y\Gamma(1/\phi))^{-1}$
$\mu(\theta)$	θ	e^θ	$-1/\theta$
linkkifunktio	μ	$\log(\mu)$	$1/\mu$

Lause 4.1. *Oletetaan, että satunnaismuuttujan Y jakauma kuuluu eksponentiaaliseen jakaumaperheeseen. Silloin satunnaismuuttujan Y:n momentti- ja kumulanttifunktio ovat*

$$M_Y(s) = \exp\left(\frac{b(\theta + \phi s) - b(\theta)}{\phi}\right)$$
$$C_Y(s) = \frac{b(\theta + \phi s) - b(\theta)}{\phi}.$$

kaikilla $s \in \mathbb{R}$ joilla $\theta + \phi s \in \Theta$.

Todistus. Todistetaan jatkuvan satunnaismuuttujan Y tapaus. Koska f on tiheysfunktio, niin

$$\int_B \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) c(y, \phi) d\upsilon(y) = 1.$$

Siis

$$\int_B \exp\left(\frac{y\theta}{\phi}\right) c(y, \phi)\upsilon(y) = \exp\left(\frac{b(\theta)}{\phi}\right),$$

jolloin integraali on olemassa kaikilla parametrien $(\theta, \phi) \in \Theta \times \Phi$ arvoilla. Seuraava lasku osoittaa, että

$$\begin{aligned}
M_Y(s) &= \mathbb{E}(e^{sY}) \\
&= \int_B \exp(sy) \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) c(y, \phi) d\upsilon(y) \\
&= \exp\left(\frac{-b(\theta)}{\phi}\right) \int_B \exp\left(\frac{(\theta + \phi s)y}{\phi}\right) c(y, \phi) d\upsilon(y) \\
&= \exp\left(\frac{-b(\theta)}{\phi}\right) \exp\left(\frac{b(\theta + \phi s)}{\phi}\right) \\
&= \exp\left(\frac{b(\theta + \phi s) - b(\theta)}{\phi}\right)
\end{aligned}$$

kaikille $s \in \mathbb{R}$ joilla $\theta + \phi s \in \Theta$. □

Lause 4.2. *Eksponentiaalisessa jakaumaperheessä pätee*

$$E(Y) = b'(\theta)$$
$$Var(Y) = \frac{b''(\theta)}{\phi}$$

Todistus. Koska jakauman odotusarvo ja varianssi voidaan määritellä kumulantit generoivan funktion avulla, niin lasketaan kumulantit generoivien

funktioiden ensimmäinen ja toinen derivaatta. Momentit generoivan funktion $M(s)$ kaksi ensimmäistä derivaattaa ovat lauseen 4.1 perusteella

$$M'(s) = b'(\theta + \phi s) \exp\left(\frac{b(\theta + \phi s) - b(\theta)}{\phi}\right) \quad \text{ja}$$
$$M''(s) = (\phi b''(\theta + \phi s) + b'(\theta + \phi s)) \exp\left(\frac{b(\theta + \phi s) - b(\theta)}{\phi}\right),$$

jolloin saadaan odotusarvo

$$\begin{aligned}\mathbb{E}(Y) &= \frac{M'(0)}{M(0)} \\ &= \frac{b'(\theta) \exp\left((b(\theta + \phi 0) - b(\theta))/\phi\right)}{\exp\left((b(\theta + \phi 0) - b(\theta))/\phi\right)} \\ &= b'(\theta),\end{aligned}$$

ja varianssi

$$\begin{aligned}Var(Y) &= \frac{M''(0)M(0) - M'(0)^2}{M(0)^2} \\ &= M''(0) - M'(0)^2 \\ &= \phi b''(\theta) + b'(\theta)^2 - b'(\theta)^2 \\ &= \phi b''(\theta).\end{aligned}$$

□

Huomautus. Lauseen 4.1 perusteella nähdään myös, että odotusarvon ja varianssin välillä on yhteys $Var(Y) = \phi b''(b'^{-1}(\mathbb{E}(Y)))$. Tämän vuoksi funktiota $V\colon b'(\Theta) \to [0, \infty[$, jossa $V(\mu) = \phi b''(b'^{-1}(\mu))$ kaikilla $\mu \in b'(\Theta)$, kutsutaan *varianssifunktioksi*.

Voidaan osoittaa, että varianssifunktio karakterisoi kunkin eksponentiaalisen jakaumaperheen kaikkien eksponentiaalisten jakaumaperheiden joukosta. Tästä hyvänä esimerkkinä ovat *Tweedie-jakaumaperheet*, joita käsitellään alaluvussa 4.2

Esimerkki 4.2. Edellisen esimerkin 4.1 perusteella $b(\theta) = e^\theta$, $e^\theta = \mu$ ja $\phi = 1$. Nyt siis

$$\mathbb{E}(Y) = \frac{\partial}{\partial\theta} b(\theta) = \frac{\partial}{\partial\theta} e^\theta = \mu \quad \text{ja}$$
$$Var(Y) = \phi \frac{\partial^2}{\partial^2\theta} b(\theta) = \frac{\partial}{\partial\theta} b(\theta) = \mu.$$

4.2 Tweedie -jakaumaperhe

Tweedie-jakaumaperhe kuuluu eksponentiaaliseen jakaumaperheeseen. Tästä seuraa se, että eksponentiaaliselle jakaumaperheelle osoitetut lauseet pätevät myös Tweedie-jakaumaperheelle. Tweedie- jakaumaperhe sisältää diskreettejä ja jatkuvia jakaumia. Lisäksi erikoistapauksena saamme sekä diskreettejä että jatkuvia jakaumia. Soveltavassa osuudessamme käytämme Tweedie-jakaumaperheestä niin kutsuttua yhdistettyä Poisson-gamma-jakaumaa, jolla pystymme yhdistämään uskottavasti vahinkotiheyden ja keskivahingon mallintamisen ominaisuudet.

Määritelmä 4.3. Olkoon $p \in \mathbb{R}$ vakio. Eksponentiaalinen jakaumaperhe on *tweedie-jakaumaperhe*, jos sen varianssifunktio $V\colon b'(\Theta) \to [0, \infty[$ toteuttaa ehdon

$$V(\mu) = \mu^p \quad \text{kaikilla } \mu \in b'(\Theta). \tag{4.2}$$

Tweedie-jakaumia eri p:n arvoilla esitetään alla olevassa taulukossa 1.

Taulukko 2: Tweedie-malleja eri p:n arvoilla

p:n arvo	Tyyppi	Jakauma
$p < 0$	Jatkuva	-
$p = 0$	Jatkuva	normaali
$0 < p < 1$	Ei-määritelty	-
$p = 1$	Diskreetti	Poisson
$1 < p < 2$	Yhdistetty ei-negatiivinen	Tweedie, yhdistetty Poisson-Gamma
$p = 2$	Jatkuva, positiivinen	Gamma
$2 < p < 3$	Jatkuva, positiivinen	-
$p = 3$	Jatkuva, positiivinen	käänteinen normaalijakama
$p > 3$	Jatkuva, positiivinen	-

Arvoilla $0 < p < 1$ eksponentiaalista jakaumaperhettä (ts. tweedie-mallia) ei ole olemassa. Jakaumaperhe on olemassa negatiivisilla p:n arvoilla, mutta työn rajoituksien vuoksi keskitymme jakaumiin, joilla $p \geq 1$.

Lause 4.3. *Funktio $b(\theta)$ on määritelty seuraavasti, kun $p \geq 1$.*

$$b(\theta) = \begin{cases} e^{\theta}, & \text{kun } p = 1, \\ -\log(-\theta), & \text{kun } p = 2, \\ -\frac{1}{p-2}(-(p-1)\theta)^{(p-2)/(p-1)}, & \text{kun } 1 < p < 2 \text{ ja } p > 2. \end{cases} \tag{4.3}$$

Parametriavaruus M_θ on määritelty seuraavasti

$$M_\theta = \begin{cases} -\infty < \theta < \infty, & \text{kun } p = 1, \\ -\infty < \theta < 0, & \text{kun } p > 1. \end{cases} \tag{4.4}$$

Funktion $b'(\theta)$ derivaatta on määritelty paloittain seuraavasti

$$b'(\theta) = \begin{cases} e^{\theta}, & kun\ p = 1, \\ (-(p-1)\theta)^{-(p-1)}, & kun\ p > 1, \end{cases} \tag{4.5}$$

ja sen käänteisfunktio

$$b'^{-1}(\mu) = \begin{cases} \log(\mu), & kun\ p = 1, \\ -\frac{1}{p-1}\mu^{-(p-1)}, & kun\ p > 1. \end{cases} \tag{4.6}$$

Todistus. Ks. [2, s. 25]. □

4.2.1 Yhdistetty Poisson-Gamma -jakauma

Tässä alaluvussa tarkastelemme Tweedie-jakaumaa, kun $p \in (1, 2)$. Tällöin jakaumaa kutsutaan yhdistetyksi *Poisson-Gamma-jakaumaksi*. Kun oletetaan esimerkiksi, että tapahtumien esiintyvyys on Poisson-jakautunut ja tapahtumien suuruus on gamma-jakautunut, niin momentit generoiva funktio käyttäen lausetta 3.9 on

$$M_Y(s) = e^{\lambda((1-\frac{s}{\beta})^{-\alpha}-1}.$$

Lisäksi momentit generoivan funktion ja lauseen 3.7 avulla voidaan osoittaa, että

$$\mathbb{E}(Y) = \frac{\lambda\alpha}{\beta}, \quad Var(Y) = \frac{\lambda\alpha(1+\alpha)}{\beta^2}. \tag{4.7}$$

Koska $\mathbb{E}(Y) = \mu$, niin saamme jakaumaperheen odotusarvolle muodon $\lambda\alpha/\beta = \mu$ ja varianssille $\lambda\alpha(\alpha+1)/\beta^2 = \phi\mu^p$, kun $\mu, \phi > 0$. Muuttujat λ, α ja β ovat määritelty seuraavasti:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1}, \qquad \frac{1}{\beta} = \phi(p-1)\mu^{p-1}. \tag{4.8}$$

Parametrit 4.8 ovat valittu siten, että satunnaismuuttujan Y yhdistetty jakauma voidaan kirjoittaa eksponentiaalisen jakaumaperheen määritelmän muodossa 4.1. Kun $y = 0$, niin Poissonin jakauman pistetodennäköisyysfunktion perusteella $P(Y = 0) = e^{-\lambda}$, ja kun $y > 0$, niin

$$f_Y(y) = e^{-\beta y}e^{-\lambda}\sum_{n=1}^{\infty}\frac{\beta^{n\alpha}}{\Gamma(n\alpha)}y^{n\alpha-1}\frac{\lambda^n}{n!}, \tag{4.9}$$

missä $\Gamma(x)$ on *gammafunktio*. Nyt parametrien 4.8 valintojen takia $\lambda\beta^{\alpha}$ ei riipu muuttujasta μ, vaan ainoastaan parametreistä ϕ ja p. Tästä syystä myös yhtälö 4.9 riippuu muuttujista ϕ ja y. Jotta voimme esittää yhtälön 4.9 eksponentiaalisen jakaumaperheen muodossa 4.1, niin määrittelemme funktion

$c(y, \phi)$ muotoon $c(0, \phi) = 1$, ja lisäksi $-\beta = \theta/\phi$ ja $\lambda = b(\theta)/\phi$. Tällöin saamme

$$\theta = -\beta\phi = \frac{-1}{(p-1)\mu^{p-1}},\ \mu(\theta) = (-\theta(p-1))^{-1/(p-1)} \text{ ja } b(\theta) = \lambda\phi = \frac{\mu^{2-p}}{2-p}.$$

Huomautus. Poisson-jakauma ($p = 1$) ja gamma-jakauma ($p = 2$) ovat rajajakaumia yhdistetyn Poisson-Gamma-jakauman tapauksessa. Tämä nähdään helposti esimerkiksi Poisson-jakauman tapauksessa, sillä kun $p \to 1$, niin jakauma lähenee Poisson-jakaumaa parametrillä μ/ϕ. Lisätietoa yhdistetystä Poisson-Gamma-jakaumasta ja rajajakaumista löytyy esimerkiksi lähteestä [9].

4.3 Yleistetyt lineaariset mallit

Yleistetyt lineaariset mallit laajentavat normaalin lineaarisen regression eksponentiaalisen perheen jakaumille. Malli voidaan jakaa kolmeen osaan:

- *Satunnainen osa* määrittelee todennäköisyysjakauman selitettävälle muuttujalle Y. Selitettävän muuttujan $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ havainnot ovat realisaatioita keskenään riippumattomista satunnaismuuttujista, joiden jakaumat kuuluvat valittuun eksponentiaaliseen jakaumaperheeseen.

- *Systemaattinen osa* määrittelee mallin selittäjät. Systemaattisessa osassa on parametrivektori $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ ja $n \times p$ mallimatriisi X, jossa on p selittävää muuttujaa n:lle havainnolle. Parametrivektori β ja mallimatriisi X muodostavat lineaarisen prediktorin η, joka on

  $$\eta = \boldsymbol{X}\boldsymbol{\beta}.$$

- *Linkkifunktio* yhdistää mallin systemaattisen osan odotusarvoon $\mathbf{E}(Y) = \mu$ seuraavasti:

  $$g(\mu) = \boldsymbol{X}\boldsymbol{\beta} = \eta.$$

  Linkkifunktio $g\colon E \to F$ on aidosti monotoninen ja derivoituva funktio, jolla on olemassa käänteisfunktio $g^{-1}\colon F \to E$. Joukko E mielletään valitun eksponentiaalisen jakaumaperheen mahdollisten odotusarvojen joukoksi, ja joukko F mielletään usein reaalilukujen joukoksi $\mathbb{R}$.

Siis selitettävän muuttujan oletetuksi jakaumaperheeksi voidaan valita taulukon 1 mukaisesti jokin eksponentiaalinen jakaumaperhe. Lisäksi selitettävän muuttujan odotusarvon ja lineaarisen prediktorin η oletettua yhteyttä mallintamaan voidaan valita linkkifunktio g. Tämän linkkifunktion ei tarvitse olla identiteettifunktio $g(\mu) = \mu$, vaan mahdollisesti voidaan valita jokin taulukon 1 vaihtoehdoista.

Kun olemme tehneet oletukset eksponentiaalisesta jakaumaperheestä ja valinneet linkkifunktion käsiteltävälle aineistolle, niin voimme valitun mallin

pohjalta voidaan muodostaa suurimman uskottavuuden estimaatit $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \in \mathbb{R}$ parametreille $\beta_0, \beta_1, \ldots, \beta_p$ sekä jokaista havaintoyksikköä kohden estimaatti $\hat{\mu}_i := g^{-1}\left(\sum_{j=1}^{p} x_{ij}\right)$ odotusarvolle $\mu_i := \mathbb{E}(Y_i)$. Lisää suurimman uskottavuuden estimaateista ja yleistetyistä lineaarisista malleista voi lukea esimerkiksi kirjasta *Foundations of Linear and Generalized Linear Models* [3], ja erityisesti vakuutusyhtiöiden aktuaaritoimintaan liittyvää tietoutta on kirjassa *Non-Life Insurance Pricing with Generalized Linear Models* [2].

4.4 Yleistetyt lineaariset mallit vakuutusten hinnoittelussa

Määritelmä 4.4. [2, s. 2-4] Määrittelemme vakuutusten hinnoittelussa ja tässä tutkielmassa käytettäviä keskeisiä käsitteitä.

- *Tariffitekijä* on vakuutuksenottajan tai vakuutettavan asian ominaisuus, jonka perusteella vakuutussopimus hinnoitellaan. Tariffitekijä luokitellaan osajoukkoihin, ja yhtä osajoukkoa kutsutaan *tariffiluokaksi*.

- *Sopimuksen kestoksi* kutsumme ajanjaksoa, jonka sopimus on voimassa vuoden aikana. Vahinkovakuutuksessa sopimukset ovat yleensä toistaiseksi voimassaolevia, ja yhden sopimuksen maksimipituus kerrallaan on yksi vuosi, jonka jälkeen se uusitaan.

- *Vahinko* on vakuutustapahtuma, josta sopimuksenhaltija vaatii taloudellista kompensaatiota menetyksestään.

- *Vahinkotiheys* on vahinkojen lukumäärä jaettuna sopimusten kestolla. Vahinkotiheys lasketaan ennalta määriteltyjen tariffiluokkien sisällä.

- *Keskivahinko* on vahingoista maksettujen korvausten summa jaettuna vahinkojen lukumäärällä. Keskivahinko lasketaan ennalta määriteltyjen tariffiluokkien sisällä.

- *Puhdas preemio* on keskimääräinen kustannus sopimukselle per vuosi. Siis

$$\text{Puhdas preemio} = \text{Vahinkotiheys} \times \text{Keskivahinko}.$$

Määritelmä 4.5. (i) Olkoon Z satunnaismuuttuja sopimukselle sattuneelle vahinkojen lukumäärälle tai sopimukselta maksetuille korvauksille. Tällöin satunnaismuuttujaa Z kutsutaan *vastemuuttujaksi*.

(ii) Olkoon w *prioripaino*. Prioripainona toimii usein vakuutusvuosien määrä tariffisolua kohden. Jos prioripaino w on vakuutusvuosi vakuutussopimuksella, niin $w \in [0, 1]$.

(iii) Satunnaismuuttujaa $Y = Z/w$ kutsutaan *vahinkomuuttujaksi.*

Teemme seuraavaksi olettamia, jotka antavat perustan tilastolliselle mallille. Olettamat harvoin toteutuvat täysin ja usein joudutaan tekemään kompromisseja perusolettamien kanssa.

Oletus 4.1 (Riippumattomuus sopimusten suhteen)**.** Oletetaan n kappaletta vakuutussopimuksia ja olkoon Z_i vastemuuttuja, kun $i \in \{1, \ldots, n\}$. Tällöin vastemuuttujat $Z_1, Z_2, \ldots, Z_n$ ovat riippumattomia.

Oletus 4.2 (Riippumattomuus ajan suhteen)**.** Olkoon Z_i vastemuuttuja. Jokaisella ajanhetkellä i, kun $i \in \{1, \ldots, n\}$ vastemuuttujat $Z_1, Z_2, \ldots, Z_n$ ovat riippumattomia.

Oletus 4.3 (Homogeenisyys)**.** Olkoon Z_i vastemuuttuja. Olkoon Z_1 ja Z_2 vastemuuttujia samasta tariffiluokasta. Tällöin vastemuuttujilla Z_1 ja Z_2 on sama todennäköisyysjakauma.

Huomautus. Riippumattomuusehto 4.1 ei täyty välttämättä aina, sillä jos esimerkiksi kaksi saman vakuutusyhtiön ajoneuvoa aiheuttavat vahingon keskenään. Lisäksi riippumattomuus ajan suhteen 4.2 ei välttämättä toteudu, sillä vakuutuksenottajalle sattuneen vahingon jälkeen hän saattaa muuttaa käytöstään alentaen omaa vahinkoherkkyyttään.

Taulukko 3: Tariffisolujen mahdollisia arvoja

Tariffisolu	Ikä	Alue	Sopimusten kesto	Vahinkotiheys
1	1	1	62,9	0,27
2	1	2	112,6	0,062
3	1	3	113,1	0,068
4	2	1	376,6	0,019
5	2	2	9,4	0
6	2	3	70,8	0,014

Oletetaan, että vahinkoaineisto on kuten yllä olevassa taulukossa 3. Havaintoyksikköinä toimivat tariffisolut, joihin vakuutussopimukset ovat luokiteltu tariffitekijöiden eri tariffiluokkien mukaan. Selitettäviä satunnaismuuttujia $Y_1, \ldots, Y_n$ kutsutaan määritelmän 4.5 mukaisesti vahinkomuuttujiksi. Alaluvussa 4.3 tehtyjen oletusten mukaan mallin selittäjät ovat tariffitekijät ja selitettävät muuttujat ovat joko vahinkotiheys tai keskivahinko. Vahinkotiheyttä usein mallinnetaan käyttämällä Poisson-jakaumaa ja keskivahinkoa käyttämällä gamma-jakaumaa.

Vakuutusten hinnoittelussa käytetään perinteisesti multiplikatiivista mallia. Multiplikatiivisessa mallissa käytetään usein log-linkkifunktiota, jolloin linkkifunktio on kuvaus $g\colon (0, \infty) \to \mathbb{R}$, jossa $g(y) = \log(y)$ kaikilla $y \in (0, \infty)$.

Esimerkki 4.3. Taulukossa 3 on yksinkertaistettu esimerkki vahinkoaineistosta, jossa on tariffitekijöinä alue ja ikä. Aineistossa ikä on jaettu kahteen ja alue kolmeen luokkaan. Jokainen näistä luokista muodostaa oman selittävän muuttujansa, ja jotta estimaatit olisivat yksikäsitteisiä [3, s. 11-15], niin jokaisesta tariffitekijästä valitaan yksi *perusluokaksi*, josta ei muodosteta selittävää muuttujaa. Perusmuuttuja on tariffitekijä, joihin muita tariffitekijän luokkia verrataan. Tässä esimerkissä valitaan perusluokiksi ikäluokka 1 ja alue 1.

Esimerkiksi kun mallinnetaan vahinkotiheyttä taulukon 3 mukaiseen aineistoon, niin oletetaan, että

(i) vahinkotiheydet $Y_1, Y_2, \ldots, Y_6$ ovat keskenään riippumattomia ja $Y_i \sim Poi(\mu_i, 1/w_i)$ kaikilla $i \in \{1, \ldots, 6\}$.

(ii) linkkifunktiona on $g\colon (0,\infty) \to \mathbb{R}$, jossa $g(y) = \log(y)$ kaikilla $y \in (0,\infty)$, jolloin mallista saadaan multiplikatiivinen.

Nyt saadaan multiplikatiivinen malli kahdella tariffitekijällä

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j},$$

missä γ_{1i} on ikäluokka, kun $i \in \{1, 2\}$ ja γ_{2j} on alueluokka, kun $j \in \{1, 2, 3\}$. Esimerkiksi perusluokka on ikäluokassa γ_{11} ja alueluokassa γ_{21}.

Nyt malli saa seuraavat kertoimet parametreilla $\beta_0, \beta_1, \beta_2$ ja β_3:

- Vakiokerroin $\mu := e^{\beta_0}$
- Ikäluokan kertoimet: $\gamma_{11} := 1$ ja $\gamma_{12} := e^{\beta_1}$.
- Alueluokan kertoimet: $\gamma_{21} := 1$, $\gamma_{22} := e^{\beta_2}$ ja $\gamma_{23} := e^{\beta_3}$.

Multiplikatiivisessa mallissa estimaattien eksponentiaaliset kertoimet e^{β_j} ovat poikkeamia asetetusta perusluokasta. Puhutaan siis suhteellisista riskeistä tariffitekijän eri luokkien välillä. Suhteellisia riskejä kutsutaan usein riskikertoimiksi. Esimerkiksi, jos $\hat{\gamma}_{12} := e^{\hat{\beta}_1} = 1,1$, niin vahinkotiheyksien odotusarvojen estimoidaan olevan 1,1 kertaiset ikäluokkaan 1 verrattuna.

4.4.1 Yhdistetty Poisson-gamma-jakauma vakuutusten hinnoittelussa

Olkoon N_i havaittu vahinkojen lukumäärä, Z_i kokonaisvahinkomäärä ja w_i prioripaino eli vakuutusvuodet i:nnessä tariffiluokassa. Nyt $Y_i = Z_i/w_i$ on keskivahinko havaitussa luokassa. Oletamme että vahinkojen lukumäärä N_i on Poisson-jakautunut $N_i \sim Poi(\lambda_i w_i)$ ja keskivahinko on gamma-jakautunut

$Y_i \sim Gamma(\alpha, \tau)$. Nyt vahinkojen suuruuden jakauma Y_i on jatkuva ja positiivinen, ja N_i ja Y_i saavat arvon nolla todennäköisyydellä $e^{-\lambda_i w_i}$. Vahinkojen sattumisajankohdat ovat riippumattomia, jolloin ehdollinen Y_i:n jakauma[1] vahinkojen lukumäärällä N_i on gamma-jakautunut parametrillä $N_i\tau_i/w_i$, kun N_i on positiivinen.

Tariffiluokkien odotusarvoksi saadaan

$$\mu_i = \mathbb{E}(Y_i) = \lambda_i \tau_i. \tag{4.10}$$

Nyt yhdistetyn jakauman varianssifunktioksi saadaan tweedie-jakauman varianssifunktion 4.7 mukaisesti

$$Var(Y_i) = \frac{\phi \mu^p}{w_i}, \tag{4.11}$$

missä parametrit p ja ϕ ovat määritelty kuten esityksissä 4.8. Parametrin p arvo estimoidaan suurimman uskottavuuden menetelmällä, joka sivuutetaan tässä tutkielmassa. Lisää tästä voi lukea esimerkiksi lähteestä [11, Jorgensen, Souza].

[1]Ehdollisen jakauman käsitettä emme käsittele tässä tutkielmassa, vaan lisätietoa voi lukea esimerkiksi lähteestä[1, Williams].

5 Sovellus: Liikennevakuutuksen hinnoittelu pakettiautoille

5.1 Aineisto ja mallin valinta

Tutkielmassa käytämme Suomen Vahinkovakuutukselta saatua aineistoa pakettiautoille sattuneista vahingoista vuosilta 2012-2016. Käytämme tutkielmassa tilastollisena ohjelmistona *RStudiota*. Selitettävänä muuttujana aineistossa on puhdas preemio, jonka oletetaan noudettavan yhdistettyä Poisson-gamma-jakaumaa luvun 4.2.1 mukaisesti. Varianssifunktion 4.2 p:n arvo tullaan estimoimaan käyttäen *tweedie.profile* funktiota, joka löytyy kirjastosta *tweedie*. Selitettävät muuttujat muodostetaan luokitelluista tariffitekijöistä, joita tässä tutkielmassa on vakuutuksenottajan ikä, ajoneuvon ikä ja vakuutuksenottajan asuinkunta. Vakuutuksenottajan ikä ja ajoneuvon ikä ovat kokonaisia vuosia. Painokertoimena mallissa on vakuutusvuosi eli sopimuksen kesto vakuutussopimuksella.

Tariffitekijät ovat luokiteltu seuraavasti:

ikäluokat: 18-24, 25-39, 40-99

ajoneuvon ikä: 0-4, 5-9, 10-14, 15-99

kuntaluokka: 1, 2, 3

Ylläolevassa tariffitekijöiden luokittelussa kuntatekijä on jaoteltu karkeasti siten, että luokassa 1 ovat pienet ja keskisuuret kunnat, luokassa 2 on keskisuuret ja suuret kaupungit, ja luokassa 3 on pääkaupunkiseutu.

Alla olevissa taulukoissa on jaoteltu tariffitekijät vakuutusvuosittain.

Taulukko 4: Vakuutuksenottajan ikäluokat

Vakuutuksenottajan ikäluokka	Vakuutusvuodet
18-24	1060,668
25-39	982,767
40-99	1397,372

Taulukko 5: Ajoneuvon ikäluokat

Ajoneuvon ikäluokka	Vakuutusvuodet
0-4	121,652
5-9	508,315
10-15	1038,611
15-99	1794,466

Taulukko 6: Kuntaluokat

Kuntaluokat	Vakuutusvuodet
1	1003,586
2	1379,370
3	1071,214

Yleisen käytännön mukaan esimerkin 4.3 mukaisiksi perusluokiksi valitseme luokat suurimman prioripainon, eli vakuutusvuosien mukaan. Toisin sanoen luokaksi valitsemme luokat, joissa on eniten havaintoja. Peruluokat ovat siis vakuutuksenottajan ikäluokka $40-99$, ajoneuvon ikäluokka $15-99$ ja kuntaluokka 2. Esitämme vielä otteen taulukosta esimerkin 3 mukaisesti, jossa tariffitekijät ovat jaoteltu soluihin

Taulukko 7: Tariffisolut

Ikäluokka	Auton ikäl.	Kuntaluokka	Vakuutusvuodet	Vahinkotiheys
18-24	0-4	1	0,9726	0,00
18-24	0-4	2	1,7123	0,00
18-24	0-4	3	1,6575	0,00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
25-39	10-14	1	60,1506	0,06649
25-39	10-14	2	110,3013	0,0362
25-39	10-14	3	129,1725	0,0928
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
40-99	15-99	1	192,8328	0,0051
40-99	15-99	2	220,8164	0,0317
40-99	15-99	3	185,4328	0,0377

Oletamme nyt, että mallissa vahinkotiheys on Poisson-jakautunut, ja vahinkojen suuruudet [2] gamma-jakautuneet. Koska tavoitteenamme on multiplikatiivinen malli, niin valitsemme linkkifunktioksi funktion g siten, että

$$g\colon (0,\infty) \to \mathbb{R}, \quad \text{jossa } g(y) = \log(y),$$

kaikilla $y \in (0,\infty)$. Valituilla perusluokilla ja linkkifunktiolla g saadaan tilastollinen malli seuraavasti:

[2] Vahinkojen suuruuksia emme esitä taulukoissa, sillä tätä tietoa emme voi paljastaa yrityksen antamasta aineistosta.

Olkoon nyt valittu malli parametreilla β_i, kun $i = \{0, 1, \ldots, 7\}$, jolloin

- Vakiokerroin $\mu := e^{\beta_0}$
- Vakuutuksenottajan ikäluokan kertoimet: $\gamma_{11} := e^{\beta_1}$, $\gamma_{12} := e^{\beta_2}$ ja $\gamma_{13} := 1$.
- Ajoneuvon ikäluokan kertoimet: $\gamma_{21} := e^{\beta_3}$, $\gamma_{22} := e^{\beta_4}$, $\gamma_{23} := e^{\beta_5}$ ja $\gamma_{24} := 1$.
- Kuntaluokan kertoimet: $\gamma_{31} := e^{\beta_6}$, $\gamma_{32} := 1$ ja $\gamma_{33} := e^{\beta_7}$.

Multiplikatiivinen malli puhtaalle preemiolle saa nyt muodon

$$\mu_{ijk} = \gamma_0 \gamma_{1i} \gamma_{2j} \gamma_{3k}, \tag{5.1}$$

missä $i, k \in \{1, 2, 3\}$ ja $j \in \{1, 2, 3, 4\}$.

5.2 Estimointi

Olemme olettaneet, että vahinkotiheys noudattaa Poisson-jakaumaa ja keskivahinko gamma-jakaumaa, joten käytämme tweedie-jakaumaperheen yhdistettyä Poisson-gamma-jakaumaa. Tällöin jakaumaperheen odotusarvo on yhtälön 4.10 mukaisesti

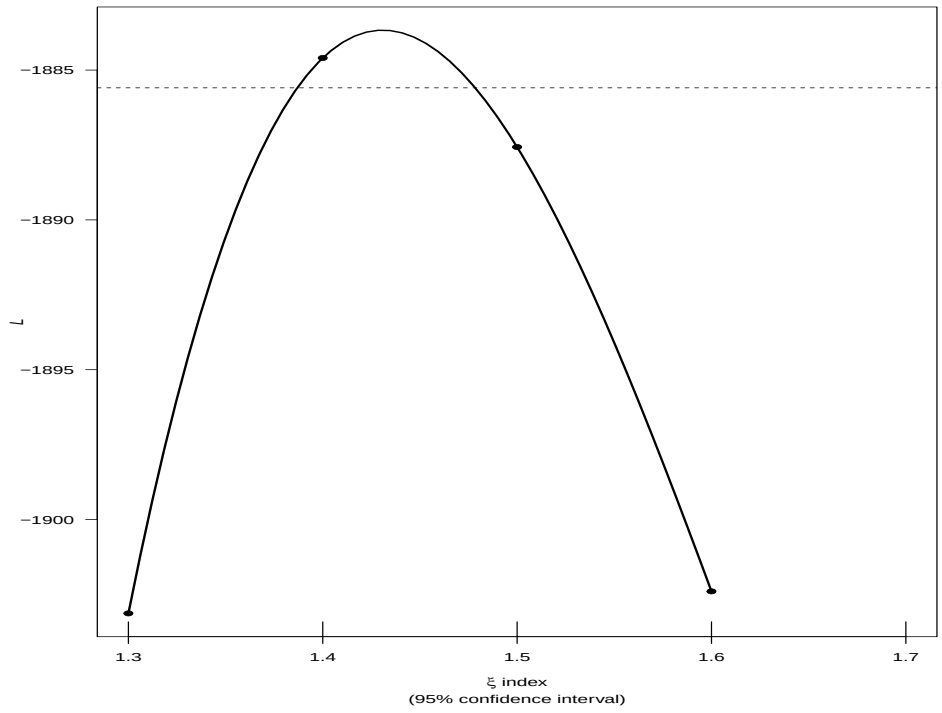
$$\mu_i = \mathbb{E}(Y_i) = \lambda_i \tau_i.$$

ja varianssifunktio on yhtälön 4.11 mukaisesti

$$Var(Y_i) = \frac{\phi \mu^p}{w_i} \quad \text{kun } p \in (1, 2).$$

Koska parametrin p arvo on tuntematon, se estimoidaan suurimman uskottavuuden menetelmällä [11] käyttäen *RStudion* tweedie-kirjaston *tweedie.profile* funktiota. Parametrin p estimoinnissa käytämme luvussa 5.1 valittua multiplikatiivista mallia. Estimoinnin tuloksena parametri p saa arvon $p = 1,4286$.

Kuva 1: Suurimman uskottavuuden arvot parametrille p



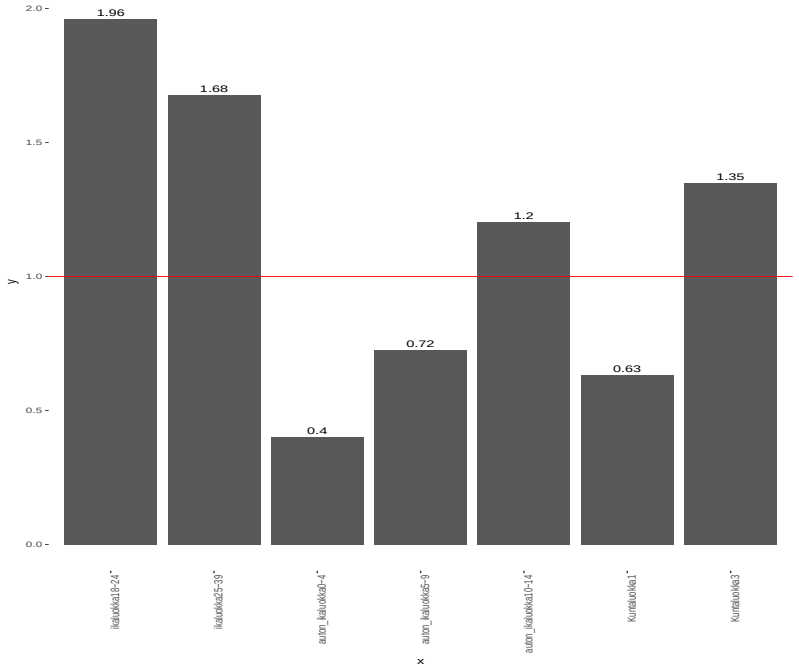
Kun estimoimme yleistetyillä lineaarisilla malleilla luvussa 5.1 valitun mallin ja käytämme *tweedie.profile* funktion antamaa parametrin $p = 1,4286$ arvoa, niin saamme seuraavat tulokset

Taulukko 8: Estimoidut arvot puhtaalle preemiolle

Tariffiluokka	Log-est.	Keskihajonta	95% :n luottamusv.	p-arvo
Perusluokka (vakio)	4,8073	0,3518	(4,097; 5,503)	<0,0001
ikaluokka 18-24	0,6729	0,3546	(-0,0360;1,385)	0,0578
ikaluokka 25-39	0,5163	0,3626	(-0,201 1,233)	0,1546
auton ikal. 0-4	-0,9130	1,0773	(-3,066; 1,056)	0,06649
auton ikal. 5-9	-0,3216	0,4756	(-1,294; 0,619)	0,0362
auton ikal. 10-14	0,1837	0,3208	(-0,459; 0,822)	0,0928
kuntaluokka 1	-0,4608	0,3749	(-1,200; 0,270)	0,0051
kuntaluokka 3	0,2995	0,3300	(-0,346; 0,942)	0,0317

Taulukossa 10 estimoidut arvot $\hat{\beta}_i$, kun $i = \{0, 1, \ldots, 7\}$ ovat niin sanotusti log-estimaatteja, jolloin esimerkiksi vakiokerroin on muotoa $\log(\mu) = \beta_0$. Teemme muunnoksen, jolloin malli saa yhtälön 5.1 mukaisen muodon. Alla olevassa taulukossa esitämme parametrien e^{β_i} estimaatit $e^{\hat{\beta}_i}$, kun $i = \{1, \ldots, 7\}$.

Kuva 2: Parametrien $e^{\hat{\beta}_i}$ estimaatit



Koska estimaatit $e^{\hat{\beta}_i}$ ovat riskikertoimia eri tariffiluokkien välillä, niin yllä olevasta taulukosta huomataan, että esimerkiksi kuntaluokassa 3 riskikertoimen odotusarvon estimoidaan olevan 1,35 kertainen perusluokkaan eli kuntaluokkaan 2 verrattuna. Vertailun vuoksi estimoimme myös erikseen va-

hinkotiheyden ja keskivahingon mallit. Vahinkotiheyttä estimoitiin Poisson-jakaumalla ja keskivahinkoa gamma-jakaumalla. Linkkifunktiona käytämme edelleen log-linkkiä. Alla ovat taulukot estimoinnin tuloksista.

Taulukko 9: Estimoidut arvot vahinkotiheydelle

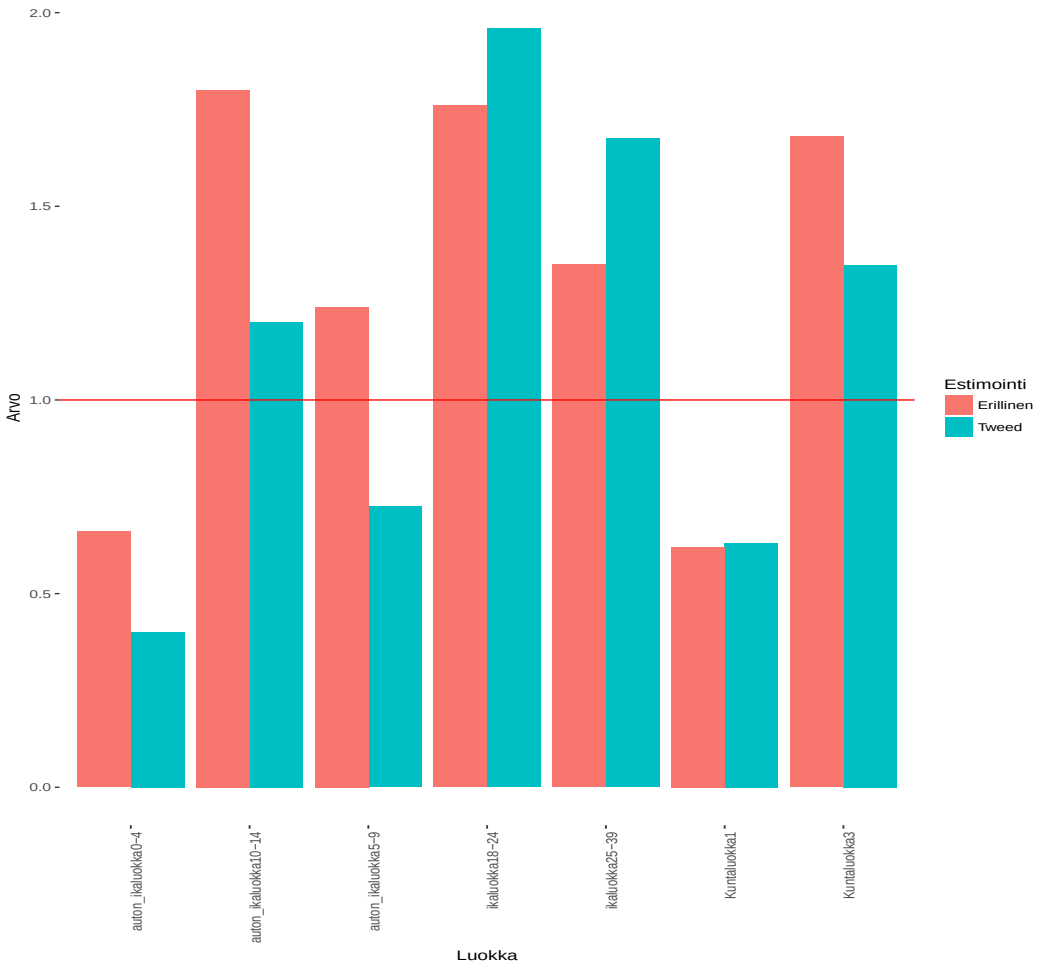
Tariffiluokka	Log-est.	Keskihajonta	95% :n luottamusv.	p-arvo
Perusluokka (vakio)	-3,5490	0,2101	(-3,975; -3,151)	<0,0001
ikaluokka 18-24	0,3665	0,2015	(-0,029; 0,763)	0,0689
ikaluokka 25-39	0,0946	0,2112	(-0,324; 0,507)	0,6542
auton ikal. 0-4	-0,2847	0,5941	-1,695; 0,717)	0,6318
auton ikal. 5-9	0,2082	0,2523	(-0,306; 0,687)	0,4093
auton ikal. 10-14	0,4311	0,1863	(0,064; 0,796)	0,0206
kuntaluokka 1	-0,3584	0,2349	(-0,833; 0,092)	0,1271
kuntaluokka 3	0,3831	0,1879	(0,015; 0,753)	0,0415

Taulukko 10: Estimoidut arvot keskivahingolle

Tariffiluokka	Log-est.	Keskihajonta	95% :n luottamusv.	p-arvo
Perusluokka (vakio)	4,9727	0,1299	(4,719; 5,238)	<0,0001
ikaluokka 18-24	0,1963	0,1365	(-0,075; 0,469)	0,150
ikaluokka 25-39	0,2068	0,1381	(-0,066; 0,482)	0,134
auton ikal. 0-4	-0,1374	0,3388	(-0,749; 0,601)	0,685
auton ikal. 5-9	0,0096	0,1815	(-0,342; 0,386)	0,957
auton ikal. 10-14	0,1575	0,1299	(-0,098; 0,419)	0,225
kuntaluokka 1	-0,1275	0,1351	(-0,391; 0,140)	0,345
kuntaluokka 3	0,1363	0,1357	(-0,127; 0,404)	0,315

Lopuksi esitämme taulukon, missä on puhtaan preemion estimaatit oletetulla tweed-jakaumalla arvolla $p = 1,4286$ ja tapauksesta, missä vahinkotiheys ja keskivahinko on estimoitu erikseen, ja tulos on yhdistetty puhtaan preemion määritelmän 4.4 mukaisesti.

Kuva 3: Estimaatit $e^{\hat{\beta}_i}$ eli riskikertoimet puhtaalle preemiolle yhdistetyllä jakaumalla ja erikseen estimoituna



5.3 Johtopäätökset

Eri menetelmien vertailua haittaa aineiston vähyys. Suurin osa vahinkotiheyden ja keskivahingon estimaateista eivät ole tilastollisesti merkitseviä. Aineiston vähyydestä huolimatta suurimpana huomiona näemme, että tariffiluokissa, missä vakuutusvuosia on yli tuhat, niin estimaatit yhdistetyllä Poisson-gamma-jakaumalla käyttäytyvät samalla tavalla kuin erikseen estimoituna. Vakuutuskannan kasvaessa eri menetelmien estimaatit lähenevät toisiaan. Luonnottoman alhainen tai korkea estimaatti on seurausta tariffiluokan alhaisesta sopimusten määrästä ja vahinkojen vähyydestä. Lisäksi molempien menetelmien estimaattien 95% luottamusväli on suhteellisen suuri johtuen karkeasta tariffiluokkien jaosta edellä mainittujen syiden lisäksi.

Työn tarkoituksena oli esitellä menetelmää eikä niinkään luoda täsmällistä mallia pakettiautojen liikennevakuutukselle. Tätä varten ajoneuvoja tulisi olla vakuutettuna moninkertainen määrä luotettavaa analyysiä varten. Tulee myös huomioida, että vakuutusyhtiön toiminnassa huomioidaan viranomaismaksut ja liiketoimintakulut, kun tehdään sovitusta eri ajoneuvoluokkien hinnoittelumalleille. Tästä syystä emme käy läpi tarkemmin yleistettyjen lineaaristen mallien analyysiin keskittyvää teoriaa emmekä esittele niitä. Halutessaan lisätietoa yleistettyjen lineaaristen mallien analyisoinnista voi lukea esimerkiksi kirjoista [3] *Foundations of Linear and Generalized Linear Models*, [2]*Non-Life Insurance Pricing with Generalized Linear Models* ja [6]*The Theory of Exponential Dispersion Models and Analysis of Deviance.*

Viitteet

[1] Williams, David *Probability with Martingales*, Cambridge University Press 1914.

[2] Ohlsson, E., Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag Berlin Heidelberg 2010.

[3] Agresti, Alan *Foundations of Linear and Generalized Linear Models*, John Wiley and Sons, Inc. Hoboken, New Jersey 2015.

[4] Ash, Robert B. *Real Analysis and Probability*, Academic Press, Inc., New York 1972.

[5] Rudin, Walter *Real and Complex Analysis*, International edition, McGraw-Hill Book Co., New York 1987.

[6] Jorgensen, Bent *The Theory of Exponential Dispersion Models and Analysis of Deviance*, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Instituto de Matemática Pura e Aplicada, Rio De Janeiro 1992.

[7] Song, P.X.-K *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer-Verlag, New York 2007.

[8] Schmidt, K. *A note on the overdispersed Poisson family*, Insurance: Mathematics and Economics 30, s. 21-15, 2002.

[9] Kaas, Rob *Compound Poisson Distributions and GLM's Tweedie's Distribution*, Department of Quantitative Economics, Universiteit van Amsterdam.

[10] Billingsley, Patrick *Probability and Measure*, John Wiley and Sons, Inc, New York 1995.

[11] Jorgensen, B., Paes de Souza, M.C. *Fitting Tweedie's compound Poisson model to insurance claim data* Scandinavian Acturial Journal, s. 69-93, 1994.

Liite: R-koodi

```
library(sqldf)
library(statmod)
library(ggplot2)
library(splines)
library(gridExtra)
library(data.table)
library(stringr)

options(scipen=999)

hakemisto<-"C:/Users/j-p.piiroinen/Desktop/Gradu/Sovellus/"
setwd(hakemisto)

vahinkoaineisto<-read.csv2(paste0(hakemisto,
                "vahinkoaineisto_N1_201210-201603.csv"))

# Kuntaluokat
kuntaluokat<-read.csv2(paste0(hakemisto,"kuntaluokka.csv"))
#Postinumerot
postinumero<-read.csv2((paste0(hakemisto,"postinumero.csv")))


ikaluokat <- c(18,25, 40, 99)
auton_ikaluokat <- c(0,5,10,15, 99)

#Korjataan kuntatiedot
yhdiste1 <- merge(x = vahinkoaineisto, y = postinumero,
                  by = "Zip.Code", all.x=TRUE)
yhdiste1$Kunta<-toupper(yhdiste1$Kunta)

#Lisätään kuntaluokat
yhdiste1 <- merge(x = yhdiste1, y = kuntaluokat,
                  by = "Kuntanumero",  all.x=TRUE)
class(yhdiste1$Kuntaluokka)
yhdiste1$Kuntaluokka<-as.factor(yhdiste1$Kuntaluokka)

#Korjataan Inception.Date faktorista päivämääräksi
yhdiste1$Inception.Date<-as.Date(
             as.character(yhdiste1$Inception.Date),
             format="%d.%m.%Y")
```

```
#Korjataan maksetut korvaukset positiiviseksi
yhdiste1$ScalTransAmtSum_MTPL<-abs(yhdiste1$ScalTransAmtSum_MTPL)

############

#Ottaa parametrinä kaksi päivämäärää
#ja laskee niiden väliset kokonaiset vuodet
age_calc <- function(startdate, enddate){
  start <- as.POSIXlt(startdate)
  end <- as.POSIXlt(enddate)
  result <- ifelse((end$mon < start$mon)|((end$mon == start$mon)
          & (end$mday < start$mday)), end$year - start$year - 1,
          end$year - start$year)
  result[result==-1] <- 0
  return(result)
}


#Ottaa parametrinä [vuosiluvun] ja
#tekee siitä päivämäärän 1.1.[vuosiluku]
vuoteen_pvm <- function(vuosi){
  pvm <- as.POSIXlt("2000-1-1")
  pvm$mday <- 1
  pvm$mon <- (1 - 1)
  pvm$year <- (vuosi - 1900)
  return(as.Date(pvm))
}

#Luodaan vakuutuksenottajan ikä ja ajoneuvon ikä -muuttujat

yhdiste1$ika <- age_calc(vuoteen_pvm(yhdiste1$YOB),
                         yhdiste1$Inception.Date)
yhdiste1$auton_ika <- age_calc(
                         vuoteen_pvm(yhdiste1$Year.of.First.Usage),
                         yhdiste1$Inception.Date)

####

cut_right = FALSE

# Luodaan luokat henkilön ja ajoneuvon iälle

yhdiste1$ikaluokka <- cut(yhdiste1$ika, breaks=ikaluokat,
                          right=cut_right)
```

```
yhdiste1$auton_ikaluokka <- cut(yhdiste1$auton_ika,
                          breaks=auton_ikaluokat, right=cut_right)

#Vakuutusvuosien taulut

ikaluokka_vv <-na.omit(sqldf('SELECT ikaluokka,
                  SUM(PolicyYears) AS PolicyYearsSum FROM yhdiste1
                  GROUP BY ikaluokka'))
auton_ikaluokka_vv <-na.omit(sqldf('SELECT auton_ikaluokka,
                  SUM(PolicyYears) AS PolicyYearsSum FROM yhdiste1
                  GROUP BY auton_ikaluokka'))
kuntaluokka_vv <-na.omit(sqldf('SELECT Kuntaluokka,
                  SUM(PolicyYears) AS PolicyYearsSum FROM yhdiste1
                  GROUP BY Kuntaluokka'))

#Tariffisolujen taulut

ryhma1 <-na.omit(sqldf('SELECT ikaluokka, auton_ikaluokka,
              Kuntaluokka, SUM(PolicyYears) AS PolicyYearsSum,
              SUM(ClaimCount_MTPL)/SUM(PolicyYears)
              AS ClaimFrequency,
              SUM(ScalTransAmtSum_MTPL) AS Severity FROM yhdiste1
              GROUP BY ikaluokka, auton_ikaluokka, Kuntaluokka'))
ryhma1
print(ryhma1, digits=2)

head(ikaluokka_vv)
head(auton_ikaluokka_vv)
head(kuntaluokka_vv)

ikaluokka_vv <- ikaluokka_vv[order(as.numeric(
   gsub("[^0-9]","", ikaluokka_vv$ikaluokka, ""))),]
auton_ikaluokka_vv <- auton_ikaluokka_vv[order(as.numeric(
   gsub("[^0-9]","", auton_ikaluokka_vv$auton_ikaluokka, ""))),]
kuntaluokka_vv <- kuntaluokka_vv[order(as.numeric(
   gsub("[^0-9]", "", kuntaluokka_vv$Kuntaluokka, ""))),]


# Etsitäään maksimit vakuutettujen vuosien mukaan
# verrokkiryhmää varten

max_ikaluokka <- ikaluokka_vv
  [which.max(ikaluokka_vv[,"PolicyYearsSum"]),]$ikaluokka
max_auton_ikaluokka <- auton_ikaluokka_vv
```

```
  [which.max(auton_ikaluokka_vv[,"PolicyYearsSum"]),]$auton_ikaluokka
max_kuntaluokka <- kuntaluokka_vv
  [which.max(kuntaluokka_vv[,"PolicyYearsSum"]),]$Kuntaluokka

yhdiste1$ikaluokka <- factor(yhdiste1$ikaluokka)
yhdiste1$ikaluokka <- relevel(yhdiste1$ikaluokka,
                              ref = max_ikaluokka)

yhdiste1$auton_ikaluokka <- factor(yhdiste1$auton_ikaluokka)
yhdiste1$auton_ikaluokka <- relevel(yhdiste1$auton_ikaluokka,
                              ref = max_auton_ikaluokka)

yhdiste1$Kuntaluokka <- relevel(yhdiste1$Kuntaluokka,
                              ref = max_kuntaluokka)

yhdiste1$PolicyYears<-yhdiste1$PolicyYears+0.000001

levels(yhdiste1$auton_ikaluokka)
      [(levels(yhdiste1$auton_ikaluokka) == "[0,5)")]<-"0-4"
levels(yhdiste1$auton_ikaluokka)
      [(levels(yhdiste1$auton_ikaluokka) == "[5,10)")]<-"5-9"
levels(yhdiste1$auton_ikaluokka)
      [(levels(yhdiste1$auton_ikaluokka) == "[10,15)")]<-"10-14"
levels(yhdiste1$auton_ikaluokka)
      [(levels(yhdiste1$auton_ikaluokka) == "[15,99)")]<-"15-99"

levels(yhdiste1$ikaluokka)
      [(levels(yhdiste1$ikaluokka) == "[18,25)")] <- "18-24"
levels(yhdiste1$ikaluokka)
      [(levels(yhdiste1$ikaluokka) == "[25,40)")] <- "25-39"
levels(yhdiste1$ikaluokka)
      [(levels(yhdiste1$ikaluokka) == "[40,99)")] <- "40-99"


#####

attach(yhdiste1)

# ScalTransAmtSum_MTPL kuvaa liikennevakuutuksesta
# maksettuja euroja

library(tweedie)
out <- tweedie.profile(ScalTransAmtSum_MTPL ~ ikaluokka
   + auton_ikaluokka + Kuntaluokka + offset(log(PolicyYears)),
```

```
   xi.vec=seq(1.3, 1.7, length=5), do.plot=TRUE, data = yhdiste1)
out$xi.max #1.428571 Liikenne

#Log-linkki ja varianssifunktio tweedie-profilesta
fit_liikenne <- glm(ScalTransAmtSum_MTPL ~ ikaluokka
   + auton_ikaluokka + Kuntaluokka + offset(log(PolicyYears)),
   data=yhdiste1,
   family=tweedie(var.power=out$xi.max, link.power=0))
fit2 <- fit_liikenne
summary(fit_liikenne)

t <- exp(coef(fit2))
tt <- t[!(names(t) %in% c('(Intercept)'))] #Poistetaan vakiotermi
ttt<-data.frame(x=names(tt), y=tt)
ttt$x <- as.character(ttt$x)
ttt$x <- factor(ttt$x, levels=unique(ttt$x))
testi<-ggplot(ttt, aes(x=x, y=y)) +
  geom_bar(stat="identity") +
  geom_text(aes(x=x, y=y, ymax=y, label=round(y,2),
  hjust=0.5, vjust=-0.5)) +
  geom_hline(yintercept=1, color="red") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_y_continuous(minor_breaks = seq(0 , 3, 0.1),
  breaks = seq(0, 3, 0.5))

confint(fit2) # 95% CI for the coefficients


testi + theme(panel.background =
  element_rect(fill = 'white', colour = 'white'))

####

yhdiste1$ScalTransAmtSum_MTPL<-yhdiste1$ScalTransAmtSum_MTPL+0.01
yhdiste1$ClaimCount_MTPL<-yhdiste1$ClaimCount_MTPL+0.0001

#Log-linkki ja Poisson
fit_liikenne1 <- glm(ClaimCount_MTPL ~ ikaluokka
  + auton_ikaluokka + Kuntaluokka + offset(log(PolicyYears)),
  data=yhdiste1, family=poisson)
fit2.1 <- fit_liikenne1
summary(fit_liikenne1)

confint(fit2.1)
```

```
#Log-linkki ja Gamma
fit_liikenne2 <- glm(ScalTransAmtSum_MTPL ~ ikaluokka
 + auton_ikaluokka + Kuntaluokka + offset(log(ClaimCount_MTPL)),
 data=yhdiste1, family=Gamma(link="log"))
fit2.2 <- fit_liikenne2
summary(fit_liikenne2)
confint(fit2.2)

a1<-c("ikaluokka18-24","ikaluokka25-39","auton_ikaluokka0-4",
 "auton_ikaluokka5-9","auton_ikaluokka10-14","Kuntaluokka1",
 "Kuntaluokka3")
a2<-c(1.9591630,1.6751814,0.4015058,0.7248419,
      1.2014232,0.6308394,1.3489770)
a2<-round(a2,2)


b1<-c("ikaluokka18-24","ikaluokka25-39","auton_ikaluokka0-4",
 "auton_ikaluokka5-9","auton_ikaluokka10-14","Kuntaluokka1",
 "Kuntaluokka3")
b2<-c(1.7556787,1.3517652,0.6556354,1.2434174,
      1.8015368,0.6151199,1.6811484)
b2<-round(b2,2)

a<-data.frame(Luokka=a1,Arvo=a2)
b<-data.frame(Luokka=b1,Arvo=b2)
a$Estimointi<-"Tweed"
b$Estimointi<-"Erillinen"

t<-rbind(a,b)

p <- ggplot(t, aes(Luokka,Arvo,fill=Estimointi))
 + geom_bar(stat="identity",position = "dodge")
 + geom_hline(yintercept=1, color="red")+
 theme(axis.text.x = element_text(angle = 90, hjust = 1))
 + geom_text(aes(x=Luokka, y=Arvo,
 label=round(Arvo,2),group=Luokka),position="dodge")


p+theme(panel.background =
 element_rect(fill = 'white', colour = 'white'))
```