Tampereen yliopisto

Walter Sirén

Todennäköisyysjakauman sovittaminen käytännön aineistoon suurimman uskottavuuden menetelmää hyödyntäen

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Huhtikuu 2020

Tiivistelmä

Walter Sirén: Todennäköisyysjakauman sovittaminen käytännön aineistoon
suurimman uskottavuuden menetelmää hyödyntäen
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastotieteen tutkinto-ohjelma
Huhtikuu 2020

Suurimman uskottavuuden menetelmä (engl. maximum likelihood estimation) on yksi suosituimmista tekniikoista todennäköisyysjakauman parametrien estimoimiseksi. Parametrien estimointi on tarpeellista esimerkiksi silloin, kun halutaan mallintaa muuttujan jakaumaa tunnetun todennäköisyysjakauman avulla. Tässä työssä suurimman uskottavuuden menetelmää hyödynnetään esimerkinomaisesti sovittamalla estimoitujen parametrien avulla todennäköisyysjakaumia muuttujan jakauman kuvaamiseksi. Sovituksia voidaan tehdä usealla erilaisella todennäköisyysjakaumalla, joista voidaan valita tiettyjä kriteereitä vertailemalla paras malli muuttujan jakauman mallintamiseksi.

Tutkimuksen voidaan ajatella jakaantuneen kahteen osaan. Ensin syvennytään teoriaosuudessa suurimman uskottavuuden menetelmään, estimaattoreiden johtamiseen, estimaattoreiden tehokkuuden arviointiin sekä informaatiokriteereiden tulkintaan. Teoriaosuuden jälkeen käytetään esitettyjä menetelmiä muuttujien jakautumisen mallintamiseen kahden eri muuttujan avulla. Tässä sovellusosassa esitellään lisäksi jakaumasovituksen onnistuminen graafisesti sekä käytetään kolmea erilaista hyvyyden mittaria sovitteen todellisen sopivuuden arviointiin. Sovellusosan jakaumasovitukset, tunnuslukujen laskelmat ja kuvaajien piirtäminen on tehty tässä tutkimuksessa R-ohjelmointikielen avulla.

Havaintoaineiston ensimmäisen muuttujan (testosteroni) jakauman mallinnus todennäköisyysjakaumalla onnistuu kohtuullisen hyvin. Parhaaksi vaihtoehdoksi valikoitunut jakaumasovite on sekä graafisesti tarkasteluna, että hyvyyden mittareita tulkitsemalla mielekäs valinta havaintojoukon testosteronitasojen malliksi. Toinen esimerkkinä käytetty muuttuja (prostataspesifinen antigeeni, PSA) paljastuu jo tarkasteluvaiheessa jakaumaltaan hyvin epäsymmetriseksi. Tälle muuttujalle valikoitunut todennäköisyysjakaumamalli ei välttämättä ole kovin hyvä. Kirjallisuusosan ja sen soveltamisen perusteella voidaan tehdä päätelmä, että aina muuttujalle ei löydy sopivaa todennäköisyysjakaumaa. Uskottavuusfunktion ja informaatiokriteereiden arvoja vertailtaessa usein jokin vaihtoehdoista nousee toisia vaihtoehtoja paremmaksi. Kandidaattimalleista parhaiten jakaumalle istuva malli ei ole kuitenkaan välttämättä sopiva.

Avainsanat: mallintaminen, parametrien estimointi, R, uskottavuusfunktio

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

Sisältö

1 Johdanto **4**

2 R-ohjelmointiympäristöstä ja fitdist()-funktiosta **5**

3 Suurimman uskottavuuden menetelmä **6**
3.1 Uskottavuusfunktio 6
3.2 Estimaattorin johtamisesta 7
3.3 Eksponenttijakauman parametrin estimointi 7
3.4 Normaalijakauman parametrien estimointi 8
3.5 Estimaattoreiden tehokkuudesta 9

4 Informaatiokriteerit AIC ja BIC **10**
4.1 Akaike information criterion (AIC) 10
4.2 Bayesian information criterion (BIC) 10

5 Menetelmien soveltaminen käytännön aineistoon **12**
5.1 Testosteroni 12
5.2 PSA eli prostataspesifinen antigeeni 15

Lähteet **17**

1 Johdanto

Tämän työn tarkoituksena on esitellä todennäköisyysjakaumien sovittamista havaintoaineistoon käyttäen suurimman uskottavuuden menetelmää. Sopivan jakauman löytämiseksi käytetään tässä tutkielmassa R-ohjelmistoa aineiston käsittelyyn ja erityisesti sen sisältämää fitdist()-funktiota eri todennäköisyysjakaumien sopivuuksien vertailemiseksi. Edellä mainittu funktio on varsinkin tämän työn laajuuden kannalta mielekäs menetelmä aineiston muuttujien mallintamiseksi todennäköisyysjakaumien avulla, ja erityisesti vaihtoehtoisten todennäköisyysjakaumien keskinäiseen vertailuun.

Todennäköisyysjakauman sovittaminen havaintoaineistoon voi olla tilastollisessa tutkimuksessa mielekäs menetelmä. Hyvin havaintoaineistoon istuva jakauma ennustaa esimerkiksi tietyn ilmiön todennäköisyyden tapahtumista siitä kokonaisesta populaatiosta, josta otos on kerätty. Tilastollisen mallintamisen periaate ylipäätänsä usein nojaa ajatukseen siitä, että erilaisten tapahtumien todennäköisyys noudattaa yhtä tai useampaa todennäköisyysjakaumaa (Hilbe ja Robinson 2013). Näiden mahdollisesti tutkittavaan ilmiöön liittyvien todennäköisyysjakaumien parametreja estimoidaan havaintoaineistosta. Tällaisia malleja, jotka ovat määritelty todennäköisyysjakaumien ja estimoitujen parametrien avulla, kutsutaan parametrisiksi malleiksi (Hilbe ja Robinson 2013).

Luvussa kaksi on eritelty lyhyesti jakaumien sovituksessa käytettävää funktiota fitdist(). Huomattavaa on, että parametrien estimointiin on olemassa muitakin tekniikoita kuin suurimman uskottavuuden menetelmä. Tässä tutkielmassa esitelläänkin perusteellisemmin vain käytettävä tekniikka, mutta funktiossa on valittavana muitakin vaihtoehtoja.

Luku kolme pyrkii käymään läpi tämän työn kannalta oleellisimman teorian käytettäessä suurimman uskottavuuden menetelmää parametrien estimointiin. Käsiteltävinä aiheina on uskottavuusfunktion määritelmä ja estimaattoreiden johtaminen yleisellä tasolla. Lisäksi käydään esimerkinomaisesti läpi estimaattoreiden johtaminen eksponenttijakaumalle sekä normaalijakaumalle.

Luvuissa neljä ja viisi siirrytään käyttämään teoriaa käytännön aineistoon. Luku neljä esittelee vielä mallien keskinäiseen vertailuun tarkoitetut informaatiokriteerit, joidenka avulla voidaan valita vaihtoehtojen joukosta sopivin malli esittämään muuttujan jakautumista. Luvussa viisi keskitytään fitdist()-funktion palauttamiin tuloksiin sovitettavista jakaumista. Tuloslistan avulla voidaan valita kadidaattimalleista parhaiten muuttujan jakaumaa kuvaava todennäköisyysjakauma.

2 R-ohjelmointiympäristöstä ja fitdist()-funktiosta

Tutkielmassa käytettävää aineistoa käsitellään R-ohjelmistolla. Käytettävä aineisto on saatu ohjaajalta Excel-työkirjana, josta tutkittavat muuttujat on siirretty vektorimuotoon R-ympäristöön. Tutkielman luonteen vuoksi muutamia puuttuvia havaintoja ei huomioida aineistoa analysoidessa, vaan ne on jätetty kokonaan pois funktioiden toimintojen helpottamiseksi.

R-ohjelmiston ehkäpä keskeisin työkalu tämän tutkielman kannalta on sen fitdist()-funktio jakaumien sovittamiseksi haluttuun aineistoon. Funktiolle välitetään parametreina tutkittava muuttuja, aineistoon sovitettava todennäköisyysjakauma sekä käytettävä menetelmä, joka on tässä työssä suurimman uskottavuuden menetelmä.

Esimerkki 2.1. fitNormal <- fitdist(x, "norm", method = "mle")

Esimerkissä 2.1 funktio saa argumenteikseen muuttujan x, sovitettavaksi todennäköisyysjakaumaksi normaalijakauman "norm" ja käytettäväksi metodiksi suurimman uskottavuuden menetelmän "mle"(engl. maximum likelihood estimation). Funktio käyttää oletusarvoisesti suurimman uskottavuuden menetelmää, joten kutsuttaessa funktiota ei ole tässä työssä tarpeen määritellä sitä erikseen. Funktion palauttama olio tallennetaan lopuksi omaan muuttujaansa fitNormal.

R-ohjelmiston summary()-funktio on geneerinen metodi jolla voidaan tarkemmin perehtyä sovitettavaan malliin. Komennolla summary(fitNormal) saadaan esimerkissä 2.1 esitetystä mallista oleellisia tietoja sovitetusta jakaumasta. Tuloslistauksesta löytyviä tietoja ovat estimoidut parametrit, parametrien keskivirheet, logaritmoidun uskottavuusfunktion arvo, parametrien estimaateista koottu korrelaatiomatriisi sekä mallien vertailuun tarkoitetut informaatiokriteerit AIC ja BIC (Delignette-Muller ja Dutang 2015).

3 Suurimman uskottavuuden menetelmä

Suurimman uskottavuuden menetelmä on yksi suosituimmista tekniikoista estimaattoreiden johtamiselle (Casella ja Berger 2002). Menetelmän periaatteena on valita se arvo jollekin parametrille θ niin, että uskottavuusfunktion arvo parametrin suhteen maksimoituu (Rohatgi 2003). Jäljempänä on esitelty tarkemmin, kuinka parametrin estimaattori johdetaan halutun todennäköisyysjakauman pistetodennäköisyys- tai tiheysfunktiosta.

Esimerkki 3.1. Jonkin aineiston havaittujen arvojen voidaan ajatella generoituvan normaalijakaumasta. Suurimman uskottavuuden menetelmää käyttämällä saadaan estimoidut arvot normaalijakauman parametreille, odotusarvolle μ ja varianssille σ^2. Saadut estimaatit parametreille ovat uskottavimmat parametrivaihtoehdot sille todennäköisyysjakaumalle, josta havaintoaineiston ajatellaan syntyneen.

3.1 Uskottavuusfunktio

Olkoon X_1, X_2, ... , X_n satunnaisotos jakaumasta $f(x|\theta)$, jonka parametrina on θ. Oletetaan havainnot toisistaan riippumattomiksi, jolloin havaintojen yhteisjakauman pistetodennäköisyys- tai tiheysfunktio on kaikkien otoksen yksittäisten havaintojen pistetodennäköisyys- tai tiheysfunktioiden tulo.

$$f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times ... \times f(x_n|\theta) \tag{3.1}$$

Uskottavuusfunktio määritellään siten

$$L(\theta|x_1, x_2, ..., x_n) = L(\theta|x_1) \times L(\theta|x_2) \times ... \times L(\theta|x_n) \tag{3.2}$$

Usein käytännössä suurimman uskottavuuden menetelmässä maksimoidaan kuitenkin uskottavuusfunktion logaritmoitu lauseke. Logaritmoitua uskottavuusfunktiota merkitään seuraavasti:

$$l(\theta) = logL(\theta) \tag{3.3}$$

Logaritmi on aidosti monotoninen funktio, joten logaritmoitu uskottavuusfunktio ja uskottavuusfunktio saavuttavat ääriarvonsa samassa pisteessä. Monesti todennäköisyysjakaumien tilanteessa logaritmoitu uskottavuusfunktio on muodoltaan yksinkertaisempi kuin uskottavuusfunktio itsessään. Logaritmin laskusääntöjen mukaan logaritminen uskottavuusfunktio voidaan esittää summamuodossa seuraavasti:

$$l(\theta) = l(\theta|x_1) + l(\theta|x_2) + ... + l(\theta|x_n) \tag{3.4}$$

3.2 Estimaattorin johtamisesta

Parametrin θ suurimman uskottavuuden estimoitu arvo saadaan maksimoimalla logaritmoitu uskottavuusfunktio parametrin suhteen. Tämä saavutetaan asettamalla uskottavuusfunktion derivaatta nollaksi, ja ratkaisemalla syntyneestä yhtälöstä mielenkiinnon kohteena oleva parametri θ. Matemaattisesti asia voidaan ilmaista seuraavalla tavalla:

$$L'(\theta) = 0 \tag{3.5}$$

Yhtälöstä ratkaistu estimaatti parametrille θ on vasta mahdollinen vaihtoehto suurimman uskottavuuden estimaatille (Casella ja Berger 2002). Estimaatti voidaan todentaa maksimiksi tarkastamalla onko uskottavuusfunktion toinen derivaatta negatiivinen (Rohatgi 2003). Formaalisti asia voidaan muotoilla seuraavanlaisesti:

$$L''(\theta) < 0 \tag{3.6}$$

Tarkastellaan seuraavaksi esimerkinmukaisesti estimaattoreiden johtamista erilaisille todennäköisyysjakaumille.

3.3 Eksponenttijakauman parametrin estimointi

Satunnaismuuttuja noudattaa eksponenttijakaumaa Exp(λ), jos sen tiheysfunktio on muotoa

$$f(x|\lambda) = \lambda e^{-\lambda x},\ kun\ \lambda > 0\ ja\ x \geq 0. \tag{3.7}$$

Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos eksponenttijakaumasta. Uskottavuusfunktio on täten muotoa

$$L(\lambda|x_1, x_2, \ldots, x_n) = \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda x_2} \times \ldots \times \lambda e^{-\lambda x_n} \tag{3.8}$$

Joka voidaan esittää myös

$$L(\lambda|x_1, x_2, \ldots, x_n) = \lambda^n e^{-\lambda(x_1+x_2+\ldots+x_n)} \tag{3.9}$$

Aikaisemmin mainittujen tietojen mukaan voidaan uskottavuusfunktio esittää logaritmoituna uskottavuusfunktiona. Tämä yksinkertaistaa uskottavuusfunktion muotoa jatkotoimenpiteitä varten. Logaritmin laskusääntöjen mukaan kohdan 3.9 logaritmoitu uskottavuusfunktio on muotoa

$$l(\lambda|x_1, x_2, \ldots, x_n) = nlog(\lambda) - \lambda(x_1 + x_2 + \ldots + x_n) \tag{3.10}$$

Seuraavaksi asetetaan saadun logaritmoidun uskottavuusfunktion ensimmäinen derivaatta parametrin λ suhteen nollaksi, ja ratkaistaan syntyneestä yhtälöstä parametri λ.

$$n\frac{1}{\lambda} - (x_1 + x_2 + \ldots + x_n) = 0 \tag{3.11}$$

$$n\frac{1}{\lambda} = (x_1 + x_2 + \ldots + x_n) \tag{3.12}$$

$$\lambda = \frac{n}{x_1 + x_2 + \ldots + x_n} \tag{3.13}$$

Näin on johdettu eksponenttijakaumalle suurimman uskottavuuden estimaattori. Kyseinen estimaattori on siis käänteisluku otoskeskiarvosta.

3.4 Normaalijakauman parametrien estimointi

Satunnaismuuttuja noudattaa normaalijakaumaa N(μ, σ^2), jos sen tiheysfunktio on muotoa

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, kun -\infty < x < \infty \tag{3.14}$$

Olkoon X_1, X_2, ... , X_n satunnaisotos normaalijakaumasta. Uskottavuusfunktio on täten muotoa

(3.15)

$$L(\mu,\sigma^2|x_1,x_2,\ldots x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

Jolloin logaritmoitu uskottavuusfunktio sievenee muotoon

$$l(\mu,\sigma^2|x_1,x_2,\ldots x_n) = -\frac{n}{2} \times ln(2\pi) - n \times ln(\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2} \tag{3.16}$$

Normaalijakauman tilanteessa estimoidaan siis kahta parametria, odotusarvoa μ ja varianssia σ^2. Ratkaistaan ensin odotusarvoparametri derivoimalla logaritmoitu uskottavuusfunktio parametrin μ suhteen, ja asettamalla se nollaksi. Kun ratkaistaan syntyneestä yhtälöstä μ, saadaan

$$\mu = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \tag{3.17}$$

Yhtälössä 3.17 on ratkaistu suurimman uskottavuuden estimaatti parametrille μ. Normaalijakauman tapauksessa se on siis otoksen keskiarvo. Estimoidaan seuraavaksi varianssiparametri σ^2 derivoimalla logaritmoitu uskottavuusfunktio parametrin σ^2 suhteen, ja asettamalla se nollaksi. Ratkaisuksi saadaan

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2}{n} \tag{3.18}$$

Huomattakoon, että suurimman uskottavuuden menetelmä tuottaa parametrille μ harhattoman estimaatin, mutta taas parametrille σ^2 estimaatti on harhainen (esim. Rohatgi 2003). Kuitenkin menetelmällä saadut estimaatit ovat yhdessä tyhjentäviä ja tarkentuvia parametreille μ ja σ^2.

3.5 Estimaattoreiden tehokkuudesta

Estimaattorilla itsellään on varianssi, joka voidaan tulkita päättelyn tekemisessä parametrin estimaattorin tehokkuudeksi (engl. efficiency). Parametrin estimaattoriksi halutaan valita sellainen harhaton estimaattori, jonka varianssi on mahdollisimman pieni. Tällaisen minimivarianssisen estimaattorin löytyminen on hyvin toivottavaa, koska sillä on vaihtoehtoisista estimaattoreista vähiten hajontaa odotusarvonsa ympärillä. Näin mahdollisimman pieni varianssi implikoi suurempaa tarkkuutta parametrin estimaattorille (ks. esim. Eliason 1993).

Harhattoman estimaattorin varianssille on johdettavissa alaraja, joka on ikään kuin mittari estimaattorin hyvyydelle (Liski 2011). Määrittelyssä käytetään hyväksi niin kutsuttua Fisherin informaatiota

$$I(\theta) = -E[\frac{d^2 l}{d\theta^2}] \tag{3.19}$$

jossa l on logaritmoitu uskottavuusfunktio. Cramérin ja Raon lauseella voidaan näin määritellä alaraja harhattoman estimaattorin varianssille. Tiettyjen säännöllisyysehtojen vallitessa estimaattorille $\hat{\theta}$ on voimassa täten

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)} \tag{3.20}$$

4 Informaatiokriteerit AIC ja BIC

Tutkielmassa käytettävä fitdist()-funktio palauttaa tuloslistauksessa kahden informaatiokriteerin arvot. Näitä kriteerejä käytetään avuksi, kun valitaan sopivaa jakaumaa muuttujan mallintamiseksi. Tässä työssä tyydytään esittelemään kriteerien käytännön tulkinta, eikä niinkään syvennytä pohtimaan niiden luotettavuutta tai keskinäisiä eroja. Kriteerien teoriaa esitellään tarkemmin muun muassa Kitagawan ja Konishin (2008) toimesta.

4.1 Akaike information criterion (AIC)

AIC on informaatiokriteeri, joka on kehitetty arvioimaan sovitettavan mallin luotettavuutta suhteessa muihin vaihtoehtoisiin malleihin. AIC voidaan kirjoittaa seuraavassa muodossa:

$$AIC = 2K - 2ln(\hat{L})$$

Yhtälössä K on mallin estimoitujen parametrien lukumäärä ja $\hat{L}$ on datasta laskettu uskottavuusfunktion estimoitu arvo. Kriteerin laskema numeerinen arvo tukee päätöksentekoa mallin valinnassa niin, että mitä pienempi arvo on, sitä tarkempi ja sopivampi sovitettava malli on. Tulkinnassa on tärkeää ottaa huomioon, että se ei yksinään kerro mallin sopivuudesta mitään, vaan ainoastaan sen suhteellisesta kelpoisuudesta muihin kandidaattimalleihin nähden. Yksittäinen arvo kriteerille ei näin ollen kelpaa mallin sopivuuden arviointiin.

Mallin parametrien määrän kasvaessa usein myös uskottavuusfunktion arvo kasvaa. AIC ottaa huomioon tämän rankaisemalla mallin ylisovittamisesta, sillä mallissa käytetyt parametrit ovat estimoituja arvoja, jotka itsessään sisältävät epävarmuutta ja virhetekijöitä.

4.2 Bayesian information criterion (BIC)

BIC on samankaltainen kriteeri kuin AIC, ja myös sitä käytetään vaihtoehtoisten mallien keskinäiseen vertailuun ja arviointiin. Myös BIC-kriteerin arvoa tulkitaan eri mallien välillä, joten sekään ei yksittäisenä arvona sisällä informaatiota mallin sopivuudesta. Aivan kuten käytettäessä AIC-kriteeriä, myös BIC tulkitaan niin, että pienempi arvo on parempi malleja vertaillessa. BIC voidaan esittää seuraavassa muodossa:

$$BIC = ln(n)K - 2ln(\hat{L})$$

Yhtälössä n on havaintojen lukumäärä, K on mallin estimoitujen parametrien lukumäärä ja $\hat{L}$ on aineistosta laskettu uskottavuusfunktion arvo. BIC siis rankaisee kompleksisia, enemmän parametreja sisältäviä, malleja AIC-kriteeriä enemmän, sillä $\ln(n) > 2$, kun $n \geq 8$.

Huomautus. Vertaillessa samalle muuttujalle sovitettuja malleja, joilla on sama määrä parametreja, ei siis ole varsinaisesti tarpeen vertailla edellä mainittujen informaatiokriteereiden arvoja. Mitä suurempi logaritmoidun uskottavuusfunktion arvo mallilla on, sitä pienempi on AIC- tai BIC-kriteerin arvo. Jos vertailtavilla malleilla on parametreja sama määrä, rankaisutermitkin (2K, ln(n)K) ovat yhtä suuret (Pawitan 2001).

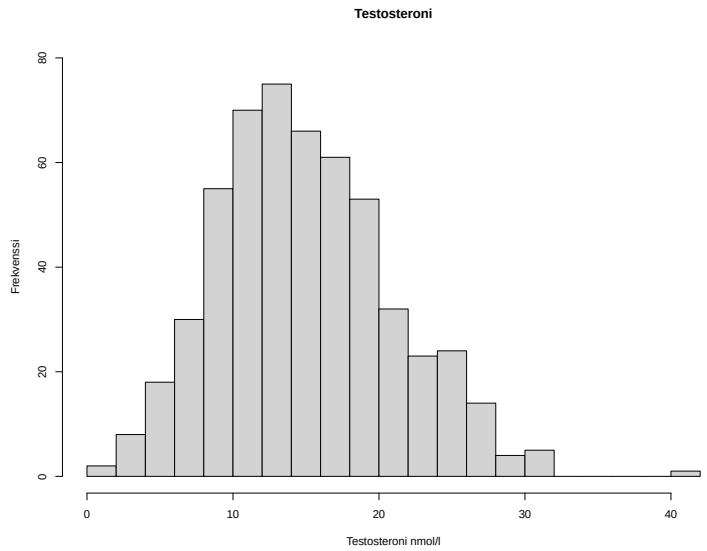
5 Menetelmien soveltaminen käytännön aineistoon

Tähän tutkielmaan on valittu käytettävästä aineistosta muutama muuttuja joita käytetään havainnollistamaan todennäköisyysjakauman sovittamista fitdist()-funktiolla. Aineisto on eturauhastutkimuksesta ja havainnot on kerätty vuosina 1997-2003. Muuttujat on eritelty omiin alalukuihinsa. Parhaan mallin löytämiseksi on jäljempänä tyydytty vain vertailemaan uskottavuusfunktion ja informaatiokriteereiden arvoja. Lisäksi R-ohjelmistolla on mahdollista piirtää kuvaajia mallin sopivuuden arvioimiseksi.

5.1 Testosteroni

Testosteroni on pääasiallinen sukuhormoni miehillä. Hormonituotannon luonnollinen heikkeneminen tapahtuu hiljalleen keski-iästä lähtien, jonka lisäksi veren testosteronipitoisuuden aleneminen voi johtua muun muuassa ylipainon kertymisestä. Erityisesti ikääntyvillä miehillä testosteronitasojen alenemisesta voi seurata yleisoireita kuten väsymystä, unettomuutta ja sukupuolista haluttomuutta. S-teston viitearvo on miehillä noin 10-38 nanomoolia litrassa (Eskelinen 2016; Mustajoki 2018).

R-ohjelmointiympäristössä voidaan käyttää funktiota hist() muuttujan jakauman esitarkastelua varten, sekä funktiota summary() muuttujan tunnuslukujen tutkimiseksi.



Kuva 5.1. Testosteronin jakautuminen histogrammilla kuvattuna.

Kuvasta 5.1 nähdään havaintojen jakauma kahteenkymmeneen luokkaan jaettuna. Havaintojoukon keskiarvo on 15.12. Ennen jakaumien sovittamista voidaan histogrammista subjektiivisesti tulkita erilaisten jakaumien järkevyyttä muuttujalle

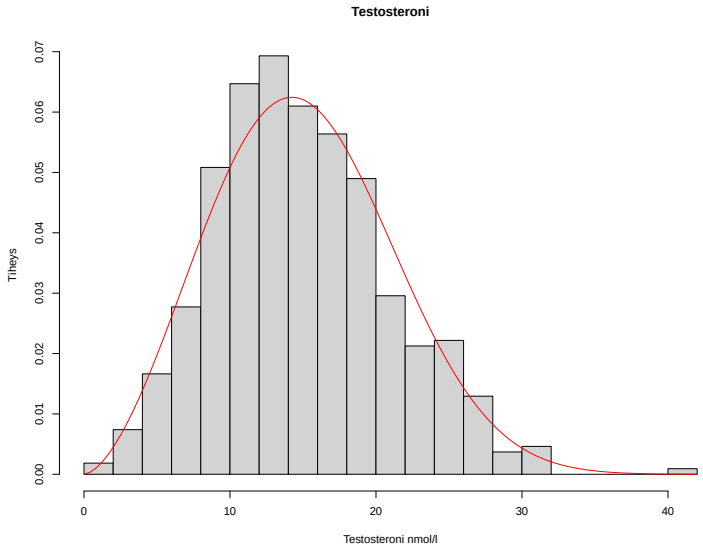
sovitettavaksi. Testosteronin jakauma on lähestulkoon symmetrinen, joten normaalijakauman sovittaminen siihen on mahdollisesti mielekäs vaihtoehto. Suurimman uskottavuuden menetelmää käyttäen saadaan kandidaattimallin parametrien estimaateiksi keskiarvolle 15.12 ja keskihajonnaksi 6.06. Logaritmisen uskottavuusfunktion arvo on -1741.97 ja informaatiokriteereiden arvot AIC:lle on 3487.94 ja BIC:lle 3496.53.

Sopivan jakauman löytämiseksi on syytä tehdä lisää sovituksia erilaisilla todennäköisyysjakaumilla. Taulukkoon 5.1 on koottu erilaisten jakaumien tietoja, joidenka perusteella parhaiten sopiva malli voidaan valita. Taulukkoon ei ole koottu estimoituja arvoja todennäköisyysjakaumien parametreille.

Taulukko 5.1. Testosteronin jakautumista kuvaavia kandidaattimalleja.

Malli	Loglik	AIC	BIC
Normaalijakauma	-1741.97	3487.94	3496.53
Gamma-jakauma	-1741.78	3487.57	3496.15
Weibull-jakauma	-1733.71	3471.42	3480.00
Log-normaalijakauma	-1775.36	3554.71	3563.30

Suurimman logaritmoidun uskottavuusfunktion arvon (-1733.71) näistä kandidaattimalleista tuottaa Weibull-jakauma. Päättelyn tekemisen tukena ovat myös informaatiokriteereiden arvot, jotka ovat Weibull-jakaumalla pienimmät. Näin voidaan todeta, että parhaiten edellä mainituista kandidaattimalleista testosteronin jakautumista kuvaa Weibull-jakauma, joka saa estimoinnin tuloksena muotoparametrin arvoksi 2.66 ja skaalausparametrin arvoksi 17.



Kuva 5.2. Testosteronin jakautuminen ja aineistosta estimoitu Weibull-jakauman malli.

Mallin hyvyyden arviointia ei voi luotettavasti perustaa pelkästään graafiseen tarkasteluun. Jakauman sopivuuden tarkasteluksi on R-ohjelmistossa tarjolla lisäksi funktio gofstat(), jonka avulla voidaan mitata etäisyyttä sovitetun ja empiirisen jakauman välillä. Etäisyyttä mitataan sovitetun jakauman kertymäfunktion F, ja empiirisen jakauman kertymäfunktion F_n välillä. Funktio gofstat() palauttaa Kolmogorov-Smirnovin (KS), Cramér-von Misesin (CvM) ja Anderson-Darlinginin (AD) hyvyyden mittarit jakaumasovituksille (Delignette-Muller ja Dutang 2015).

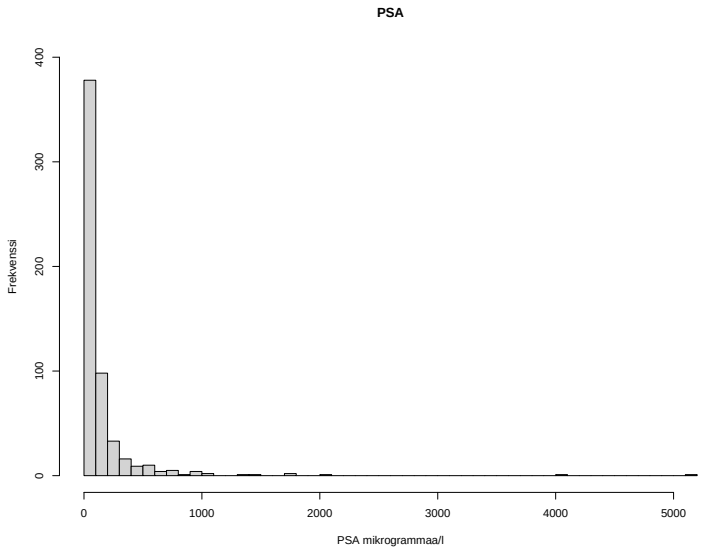
Taulukko 5.2. Mallin hyvyyden mittareita testosteronin jakaumasovituksille.

Malli	KS	CvM	AD
Normaalijakauma	0.0525	0.302	1.846
Gamma-jakauma	0.0435	0.191	1.284
Weibull-jakauma	0.0424	0.121	0.733
Log-normaalijakauma	0.0727	0.792	5.062

Taulukkoon 5.2 on koottu testosteronin jakaumasovituksien etäisyydet vertailun tekemiseksi. Jakaumasovituksien hyvyyden mittarit puoltavat myös Weibull-jakauman valintaa, koska kaikkien hyvyysmittareiden arvot ovat sillä pienimmät kandidaattimalleja vertaillessa. Tuloslistauksessa on testitulokset hypoteesille, jonka mukaan havaittu otos olisi sovitetusta jakaumasta. Esimerkiksi testattaessa hypoteesia Anderson-Darlingin hyvyysmittarilla ainoastaan Weibull-jakauman kohdalla testi ei hylkää hypoteesia. Merkitsevyystaso on funktiossa oletusarvoisesti 0.05. Mallien vertailu ei kuitenkaan ole aina suoraviivaisesti tehtävissä hyvyyden mittareiden avulla, vaan valintaa tehdessä on otettava huomioon esimerkiksi mallin parametrien lukumäärä (ks. Delignette-Muller ja Dutang 2015).

5.2 PSA eli prostataspesifinen antigeeni

PSA on eturauhasolujen tuottama valkuaisaine, jota esiintyy pienissä määrin veriplasmassa. Eturauhassyöpää sairastavalla potilaalla PSA-arvo saattaa usein nousta huomattavasti, ja sitä käytetäänkin apuna syövän toteamiseen ja hoidon vaikuttavuuden arviointiin. Iän noustessa eturauhanen yleensä kasvaa, mikä myös osaltaan nostaa PSA-arvoa. Viitearvot ovat alle 56-vuotiailla alle 3.3 mikrogrammaa litrassa ja alle 67-vuotiailla alle 6.2 mikrogrammaa litrassa (Eskelinen 2017).



Kuva 5.3. PSA:n jakautuminen histogrammilla kuvattuna.

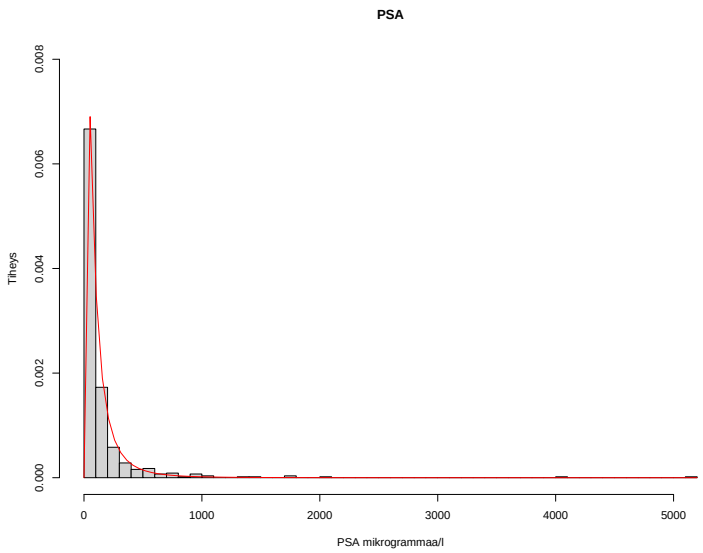
Kuvasta 5.3 nähdään PSA:n jakautuminen havaintojoukossa. Tarkastellessa jakaumaa histogrammin avulla huomataan, että muuttujan jakauma ei ole symmetrinen. Normaalijakauman käyttäminen muuttujan mallina ei siis todennäköisesti ole paras vaihtoehto, mikä huomataan myös jäljempänä esitetyissä tuloksissa. Funktiolla summary() saadaan tarkempia tunnuslukuja PSA:lle, joista mainittakoon mediaani 67.2 ja keskiarvo 150. Kolmas kvartiili on 124.5, eli 75 prosenttia havainnoista on välillä 0.9-124.5. Suurin arvo havaintojoukossa on 5123.

Taulukko 5.3. PSA:n jakautumista kuvaavia kandidaattimalleja.

Malli	Loglik	AIC	BIC
Normaalijakauma	-4115.08	8234.16	8242.84
Gamma-jakauma	-3402.01	6808.03	6816.71
Weibull-jakauma	-3378.26	6760.53	6769.21
Log-normaalijakauma	-3274.77	6553.53	6562.21
Eksponenttijakauma	-3407.96	6817.92	6822.26

Taulukkoon 5.3 on kerätty PSA:lle sovitettuja malleja. Kandidaattimalleista parhaiten PSA:n jakautumista mallintaa siis log-normaalijakauma jolla on vertailtavista

malleista suurin logaritmisen uskottavuusfunktion arvo, -3274.77. Lisäksi informaatiokriteereiden AIC:n ja BIC:n arvot ovat vaihtoehdoista pienimmät. Parametreikseen jakauma saa skaalausparametrille arvon 4.31 ja muotoparametrille arvon 1.05.



Kuva 5.4. PSA:n jakautuminen ja aineistosta estimoitu log-normaalijakaumamalli.

Tutkitaan myös hyvyysmittareiden arvoja PSA:lle sovitetuista jakaumista. Tulokset on koottu taulukkoon 5.4. Vaihtoehtoisista malleista log-normaalijakaumalla on pienimmät hyvyyssmittareiden arvot. Huomioitavaa on, että vaikkakin kandidaattimalleja vertaillessa löytyy paras vaihtoehto, ei se välttämättä ole sopiva malli. Funktion gofstat() tuloslistauksesta löytyy myös testitulokset hypoteesille, jonka mukaan havaittu otos olisi sovitetusta jakaumasta. Merkitsevyystasolla 0.05 testi hylkää hypoteesin jokaisen PSA:n jakaumasovitteen kohdalla. Log-normaalijakauma on vaihtoehdoista paras, mutta ei siltikään kovin hyvä malli.

Taulukko 5.4. Mallin hyvyyden mittareita PSA:n jakaumasovituksille.

Malli	KS	CvM	AD
Normaalijakauma	0.3340	21.895	Inf
Gamma-jakauma	0.1728	5.546	28.90
Weibull-jakauma	0.1502	3.915	23.42
Log-normaalijakauma	0.0678	0.807	5.08
Eksponenttijakauma	0.1930	7.222	34.83

Lähteet

Casella, G. & Berger, R, (2002). *Statistical inference*. Pacific Grove:Duxbury.

Delignette-Muller, M. & Dutang, C. (2015). *fitdistrplus: An R Package for Fitting Distributions*. Journal of Statistical Software, 64(4), 1-34. URL: http://www.jstatsoft.org/v64/i04/. Viitattu 7.4.2020.

Eliason, S. (1993). *Maximum likelihood estimation*. Newbury Park, Calif: Sage.

Eskelinen, S. (2017). *Prostataspesifinen antigeeni, "eturauhaskoe"(P-PSA)*. www.terveyskirjasto.fi. Viitattu 7.4.2020.

Eskelinen, S. (2016). *Testosteroni (S-testo ja S-TestoVI)*. www.terveyskirjasto.fi. Viitattu 7.4.2020.

Hilbe, J. & Robinson, A. (2013). *Methods of statistical model estimation*. Boca Raton: CRC Press.

Konishi, S. & Kitagawa, G. (2008). *Information criteria and statistical modelling*. https://doi.org/10.1007/978-0-387-71887-3

Liski, E. (2011). *Matemaattinen tilastotiede*, luentomoniste. Tampereen yliopisto.

Mustajoki, P. (2018). *Testosteroni ikääntyvillä miehillä*. www.terveyskirjasto.fi. Viitattu 7.4.2020.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford: Clarendon Press.

Rohatgi, V. (2003). *Statistical inference*. Mineola (N.Y.): Dover.