muotoa

$$\log f_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{p} \eta_i(\boldsymbol{\theta}) T_i(x) - A(\boldsymbol{\theta}) + c(x), \tag{10.9.1}$$

missä $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ ja $A(\boldsymbol{\theta}), c(x), \eta_i(\boldsymbol{\theta})$ ja $T_i(x)$ ovat tunnettuja funktioita. Lisäksi muuttujan x arvoalue ei riipu tuntemattomasta parametrista $\boldsymbol{\theta}$. Parametreja η_i kutsutaan luonnollisiksi parametreiksi ja tunnuslukuja $T_i(x)$ luonnollisiksi tunnusluvuiksi. Oletetaan lisäksi, että luonnolliset parametrit η_i eivät ole keskenään lineaarisesti riippuvia eivätkä luonnolliset tunnusluvut $T_i(x)$ keskenään. Voidaan osoittaa, että luonnolliset tunnusluvut $T_i(x)$ ovat minimaalisesti tyhjentäviä.

Eksponentiaalinen perhe on täysiasteinen, jos luonnollisten parametrien avaruus sisältää avoimen joukon. Esimerkiksi 2-ulotteinen neliö 2-ulotteisessa (2D) Euklidisessa avaruudessa sisältää avoimen joukon, mutta esimerkiksi suora 2D-avaruudessa ei sisällä avointa joukkoa. Tyypillisesti eksponentiaalinen perhe on täysiasteinen, jos luonnollisten parametrien lukumäärä on sama kuin tuntemattomien parametrien lukumäärä.

Jos luonnollisten parametrien välillä on epälineaarisia riippuvuuksia, niin luonnollisten tyhjentävien parametrien lukumäärä on suurempi kuin vapaitten parametrien lukumäärä ja silloin eksponentiaalista perhettä kutsutaan kaareutuvaksi. Monet teoreetiset tulokset pitävät paikkansa vain täysiasteisissa malleissa. Eksponentiaalinen perhe sisältää sekä jatkuvia että diskreettejä satunnaismuuttujia. Esimerkiksi normaali-, binomi- ja gammajakauma kuuluvat eksponentiaaliseen perheeseen, kun taas Cauchyn jakauma ja t-jakauma eivät kuulu.

Esimerkki 10.10 Tarkastellaan normaalijakaumaa $\mathrm{N}(\boldsymbol{\theta})$, missä $\boldsymbol{\theta} = (\mu, \sigma^2)$ ja molemmat parametrit ovat tuntemattomia. Silloin jakauman tiheysfunktion logaritmi on

$$\log f_{\boldsymbol{\theta}}(x) = \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).$$

Tämä normaalijakauma on täysiasteinen eksponentiaalisen perheen malli, luonnolliset parametrit ovat $\eta_1 = \frac{\mu}{\sigma^2}$ ja $\eta_2 = -\frac{1}{2\sigma^2}$ sekä luonnolliset tunnusluvut $T_1(x) = x$ ja $T_2(x) = x^2$.

Oletetaan nyt, että satunnaismuuttujan $X \sim \mathrm{N}(\boldsymbol{\theta})$ variaatiolerroin on tunnettu vakio $a = \sigma/\mu$, missä $\mu > 0$. Silloin $\mathrm{N}(\mu, a^2\mu^2)$ kuuluu kaareutuvaan eksponentiaaliseen perheeseen. Vaikka mallissa on vain yksi tunematon parametri, luonnolliset tunnusluvut $T_1(x) = x$ ja $T_2(x) = x^2$ ovat edelleen minimaalisesti tyhjentäviä. □

Esimerkki 10.11 Poissonin jakauman $\mathrm{Poi}(\mu)$ tapauksessa todennäköisyysfunktion logaritmi on

$$\log f_\mu(x) = x \log \mu - \mu - \log x!.$$

Siksi Poi(μ) on täysiasteinen yhden parametrin eksponentiaalisen perheen malli.

Jos X noudattaa sellaista typistettyä Poissonin jakaumaa, että arvoa $X = 0$ ei havaita, niin arvojen $x = 1, 2, \ldots$ todennäköisyydet ovat

$$f_\mu(x) = \frac{e^{-\mu}\mu^x/x!}{1 - e^{-\mu}},$$

joten

$$\log f_\mu(x) = x \log \mu - \mu - \log(1 - e^{-\mu}) - \log x!.$$

Tämä on myös yhden parametrin eksponentiaalisen perheen malli, jolla on samat kanoninen parametri ja tunnusluku, mutta funktio $A(\mu)$ on muuttunut. □

Jos $X_1, \ldots, X_n$ on otos eksponentiaalisen perheen jakaumasta, niin otoksen yhteisjakauma kuuluu eksponentiaaliseen perheeseen. Olkoon $x_1, \ldots, x_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, niin yhteisjakauman tiheysfunktion logaritmi on

$$\log f_{\boldsymbol{\theta}}(x_1, \ldots, x_n) = \frac{\mu \sum_i x_i}{\sigma^2} - \frac{\sum_i x_i^2}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2).$$

Tässä jakaumalla on samat luonnolliset parametrit $\eta_1 = \frac{\mu}{\sigma^2}$ ja $\eta_2 = -\frac{1}{2\sigma^2}$ kuin normaalijakaumalla Esimerkissä 10.10. Yhteisjakauman luonnolliset tunnusluvut ovat $T_1(x_1, \ldots, x_n) = \sum_i x_i$ ja $T_2(x_1, \ldots, x_n) = \sum_i x_i^2$.

Tarkastellaan nyt lähemmin funktion $A(\theta)$ roolia. Oletetaan, että X on satunnaismuuttuja, jonka tiheysfunktio on $\exp[c(x)]$ ja momenttifunktio

$$m(\theta) = E(e^{\theta X}).$$

Olkoon $A(\theta) \equiv \log m(\theta))$, jota kutsutaan tavallisesti kumulantteja generoivaksi funktioksi. Silloin

$$\int e^{\theta x + c(x)} d\, x = e^{A(\theta)}$$

tai

$$\int e^{\theta x - A(\theta) + c(x)} d\, x = 1$$

kaikilla θ. Näin siis funktio

$$f_\theta(x) = e^{\theta x - A(\theta) + c(x)} d\, x \tag{10.9.2}$$

määrittelee eksponentiaaliseen perheeseen kuuluvan tiheysfunktion (tai todennäköisyysfunktion). Tätä menetelmää kutsutaan satunnasimuuttujan eksponenttikäännöksi. Alkuperäinen satunnasimuuttuja X liittyy arvoon $\theta = 0$.

Jokaiselle muoroa (10.9.2) olevalle funktiolle pätee, että

$$\mu = E_\theta(X) = A'(\theta)$$

ja

$$\mathrm{Var}_\theta(X) = A''(\theta) = \frac{\partial}{\partial \theta} E\,\theta(X) = v(\mu).$$

Siksi $A(\theta)$ määrittelee tietyn yhteyden odotusarvon ja varianssin välille.

Eksponentiaalinen hajontamalli

Oletetaan nyt, että parametrien θ ja ϕ logaritmoitu uskottavuusfunktio on muotoa

$$\log L(\theta,\phi) = \frac{x\theta - A(\theta)}{\phi} + c(x,\phi), \tag{10.9.3}$$

missä $A(\theta)$ ja $c(x,\phi)$ ovat annettuja funktioita. Tässä esitysmuodossa parametria θ kutsutaan kanoniseksi parametriksi ja ϕ:tä hajontaparametriksi. Valitsemalla funktiot $A(\theta)$ ja $c(x,\phi)$ eri tavoin saadaan määriteltyä erilaisia todennäköisyysmalleja. Jos X on diskreetti, niin silloin täytyy olla

$$\sum_x \exp\left\{\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right\} = 1,$$

mikä määrittelee $A(\theta)$:n ja $c(x,\phi)$:n välille tietyn yhteyden.

Hajontaparametri sallii varianssin vaihdella vapaasti odotusarvosta riippumatta:

$$\mu = E_\theta(X) = A'(\theta)$$

ja

$$\begin{aligned}\mathrm{Var}_\theta(X) =& \phi A''(\theta)\\ =& \phi\frac{\partial}{\partial\theta} E\,\theta(X) = \phi v(\mu).\end{aligned}$$

Tämä on eksponentiaalisen hajontamallin suurin etu jäykempään malliin (10.9.1) verrattuna.

Käytännössä funktio $A(\theta)$ annetaan mallissa (10.9.3) eksplisiittisesti, kun taas funktio $c(x,\phi)$ jätetään tavallisesti tarkemmin määrittelemättä. Tämä ei ole ongelma niin kauan kuin on kyse vain θ:n estimoinnista, koska pistefunktio ei riipu $c(x,\phi)$:sta. Sen sijaan ϕ:n estimoimiseksi tarvitaan koko uskottavuusfunktio.

Esimerkki 10.12 Normaalijakauma $\mathrm{N}(\mu,\sigma^2)$ logaritmoitu uskottavuusfunktio voidaan kirjoittaa muodossa

$$\log L(\mu,\sigma^2) = \frac{x\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{x^2}{2\sigma^2}.$$

Normaalijakauma on siis eksponentiaalisen perheen malli, jonka kanoninen parametri on μ, hajontaparametri $\phi = \sigma^2$, $A(\mu) = \mu^2/2$ ja $c(x,\phi) = -\frac{1}{2}\log\phi - \frac{1}{2}x^2/\phi$. Tässä siis $c(x,\phi)$:n lauseke tunnetaan. □

Esimerkki 10.13 Poissonin jakauman $\mathrm{Poi}(\mu)$ logaritmoitu uskottavuusfunktio on muotoa

$$\log L(\mu) = x\log\mu - \mu - logx!.$$

Tässä mallissa kanoninen parametri on $\theta = \log \mu$, hajontaparametri $\phi = 1$ ja $A(\theta) = \mu = e^\theta$.

Pitämällä $A(\theta) = e^\theta$ vakiona ja vaihtelemalla hajonataparametria ϕ voidaan muodostaa Poissonin malleja, joissa on joko yli- tai alihajontaa. Samalla odotusarvolla $E(X) = \mu$ saadaan

$$\mathrm{Var}(X) = \phi A''(\theta) = \phi\mu.$$

Funktion $c(x, \phi)$ täytyy toteuttaa identiteetti

$$\sum_{x=0}^{\infty} \exp\left[\frac{x\theta - e^\theta}{\phi} + c(x, \phi)\right] = 1,$$

kaikilla θ ja ϕ. □

Esimerkki 10.14 Tarkastettiin 20 uuden auton otoksessa värivikojen lukumäärä käyttäen tiettyä kokeellista menetelmää. Tulokseksi saatiin seuraavat lukumäärät:

`0 10 1 1 1 2 1 4 11 0 5 2 5 2 0 2 0 1 3 0`

Otoskeskiarvo on $\overline{x} = 2.55$ ja otosvarianssi $s^2 = 9.85$, mikä osoittaa ylihajontaa. Jos käytetään Poissonin tyyppistä mallia, jossa $A(\theta) = e^\theta$, niin

$$\mathrm{Var}_\theta(X) = \phi\, E_\theta(X).$$

Momenttimenetelmällä hajontaparametrin ϕ estimaatti on $\hat{\phi} = 9.84/2.55 = 3.86$. Aivan ilmeisesti hajontaparametrin arvo on suurempi kuin 1. □

Minimaalinen tyhjentävyys ja eksponentiaalinen perhe

Minimaalisen tyhjentävyyden ja eksponentiaalisen perheen välillä on tärkeä teoreettinen yhteys. Lievien oletusten vallitessa tyhjentävä (minimaalisesti tyhjentävä) tunnusluku (tunnusvektori) on olemassa jos ja vain jos $f_\theta(x)$ kuuluu täysiasteiseen eksponentiaaliseen perheeseen. Eksponentiaalisen perheen rakenteesta taas seuraa, että uskottavuusfunktiolla on yksikäsitteinen maksimi ja SUE on tyhjentävä. Jos siis tyhjentävä estimaatti (tunnusluku) on olemassa, niin se on SUE.

10.10 Boxin ja Coxin muunnosperhe

Box ja Cox (1964) esittelivät positiivisille ei-normaalisille satunnaismuuttujille käyttökelpoisen muunnosperheen. Muunnoksella pyritään saattamaan vino jakauma likimain normaaliseksi. Havaitsemme satunnaismuuttujan Y arvoja. Boxin ja Coxin muunnoksessa oletetaan, että on olemassa sellainen vakio $\lambda \neq 0$, että satunnaismuuttuja

$$Y_\lambda = \frac{Y^\lambda - 1}{\lambda}$$

noudattaa normaalijakaumaa $\mathrm{N}(\mu, \sigma^2)$. Kun $\lambda \to 0$, niin $Y_\lambda \to \log Y$. Näin ollen myös logaritmimuunnos kuuluu Boxin ja Coxin muunnosperheeseen.

Yhteen havaintoon $Y = y$ perustuva logaritmoitu uskottavuusfunktio on

$$\begin{aligned}\log L(\lambda, \mu, \sigma^2) &= \log f(y_\lambda) + (\lambda - 1)\log y \\ &= -\frac{1}{2}\log\sigma^2 - \frac{(y_\lambda - \mu)^2}{2\sigma^2} + (\lambda - 1)\log y,\end{aligned} \tag{10.10.1}$$

missä $f(y_\lambda)$ on y_λ:n tiheysfunktio ja $(\lambda - 1)\log y$ on jakobiaanin logaritmi.

On huomattava, että Y_λ ei voi noudattaa normaalijakaumaa täsmällisesti, jos Y on positiivinen. Silloin voi kokeilla typistettyä normaalijakaumaa. Jos $Y_\lambda > 0$ noudattaa typistettyä normaalijakaumaa, sen tiheysfunktio on

$$f(y_\lambda) = \frac{\sigma^{-1}\phi[(y_\lambda - \mu)/\sigma]}{1 - \Phi(-\mu/\sigma)},$$

missä $\phi(\cdot)$ on standardinormaalijakauman tiheysfunktio ja $\Phi(\cdot)$ kertymäfunktio. Typistettyyn normaalijakumaan liittyvä uskottavuusfunktio saadaan lisäämällä uskottavuusfunktioon (10.10.1) termi $\log[1 - \Phi(-\mu/\sigma)]$.

Joskus muunnos toimii paremmin, jos sitä sovelletaan siirrettyihin arvoihin eli

$$y_\lambda = \frac{(y + c)^\lambda - 1}{\lambda}.$$

Siirtoparametri c voidaan estimoida aineistosta maksimoimalla uskottavuusfunktio.

Muunnoksen jälkeen parametrien μ ja σ^2 tulkinta alkuperäisten havaintojen avulla voi tuottaa ongelmia. Koska Boxin ja Coxin muunnos on monotoninen, y_λ:n mediaani on alkuperäisen aineiston y mediaanin muunnettu arvo. Jos esimerkiksi $\lambda = 0$ ja y_λ:n mediaani on 0, silloin alkuperäisen aineisto mediaani on $e^0 = 1$. Pääasiallisin syy Boxin ja Coxin muunnoksen käyttöön on pyrkimys saada muunnetut arvot noudattamaan mahdollisimman läheisesti normaalijakaumaa. Tällainen asteikko johtaa tyypillisesti yksinkertaisempaan analyysiin.

Esimerkki 10.15 Seuraavassa on 27 Euroopan Unioniin kuuluvan maan väkiluvut:

```
82 59 59 57 39 38 22 16 10 10 10 10 10 9 9 8 5 5 5 4 4 2 2 1 0.7
0.4 0.4
```

Jakauma on oikealle vino. Profiiliuskottavuus maksimoituu, kun $\hat{\lambda} = 0.12$, mikä osoittaa, että muunnos $\lambda = 0$ on uskottava vaihtoehto. □

10.11 Sijainti-skaala jakaumaperhe

Sijainti-skaala perheen jakaumat voidaan esittää parametrien μ ja σ sekä tunnetun tiheysfunktion $f_0(\cdot)$ avulla siten, että jokaisen perheeseen kuuluvan jakauman tiheysfunktio voidaan esittää muodossa

$$f(x) = \frac{1}{\sigma} f_0(\frac{x-\mu}{\sigma}).$$

Parametri μ on sijaintiparametri, σ skaalaparametri ja $f_0(\cdot)$ on perheen standardimuotoinen tiheysfunktio. Esimerkiksi normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ tapauksessa standarditiheys on

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Toinen tunettu esimerkki on Cauchy(μ, σ), jonka standarditiheys on

$$f_0(x) = \frac{1}{\pi(1+x^2)}.$$

Cauchyn jakauma on ns. paksuhäntäinen jakauma, jossa jakauman häntien alla on normaalijakaumaa enemmän todennäköisyysmassaa.

Esimerkki 10.16 Kun merkittiin muistiin kahden kahden paikkakunnan aurinkosäteilyn maksimien erotukset tietyltä ajanjaksolta, saatiin seuraava aineisto (havainnot suuruusjärjestyksessä):
`-26.8 -3.5 -3.4 -1.2 0.4 1.3 2.3 2.7 3.0 3.2 3.2 3.5 3.6 3.9 4.2`
`4.4 5.0 6.5 6.7 7.1 8.1 10.5 10.7 24.0 32.8`
Kun aineistosta tehdään normaalijakaumakuvio, huomataan aineiston paksut hännät. Aineiston keskiarvo on 3.6 ja mediaani 4.5. Cauchyn jakauman Cauchy(μ, σ) uskottavuusfunktio on

$$L(\mu, \sigma) = \prod_i \frac{1}{\sigma} \left[1 + \frac{x_i - \mu}{\sigma^2} \right],$$

josta voidaan laskea μ:n profiiliuskottavuus. Parametrin σ SUE, kun μ kiinnitetty, pitää määrittää numeerisesti, koska sillä ei ole suljetun muodon ratkaisua. Cauchyn jakaumasta saadaan SUE:t $\hat{\mu} = 3.7$ ja $\hat{\sigma} = 2.2$. □

Voimme myös valita esimerkiksi Akaiken informaatiokriteerillä AIC joko nomaalijakauman tai Cauchyn jakauman. Säteilyaineistosta normaalijakauman AIC on

$$\begin{aligned} AIC =& -2 \log L(\hat{\mu}, \hat{\sigma}^2) + 4 \\ =& n \log(2\pi\hat{\sigma}^2) + n + 4 = 189.7, \end{aligned}$$

missä $n = 25$ ja $\hat{\sigma}^2 = 98.64$. Cauchyn jakaumaa käyttäen $AIC = 169.3$, joten myös AIC:n nojalla Cauchyn malli on parempi.