Luku 7

Johdanto tilastolliseen päättelyyn

7.1 Tilastollisen ongelman luonne

Tilastollisen mallintaminen ja päättely käsittelee vaihtelua ja epävarmuutta. Ei ole itsestään selvää, että tuollaisista aiheista voidaan esittää jotain täsmällistä tai tieteellistä. Tilastotieteessäkin on omaksuttu erilaisia lähestymistapoja epävarmuuden käsittelyyn. Kaksi pääkoulukuntaa ovat Bayesilainen ja frekventistinen koulukunta. Seuraava esitys perustuu pääoson uskottavuuden (uskottavuusfunktio) käsitteeseen, joka on kummankin edellä mainitun koulukunnan keskeinen peruskäsite .

Aspiriiniaineisto

Taulukon 7.1 tulokset ovat peräisin tutkimuksesta, jossa selvitettiin, ehkäiseekö aspiriini aivohalvauksia ja sydäninfarkteja (infarctus myocardii acutus) (Steering Committee of the Physicians' Health Study Research Group 1989). Tutkimuksessa satunnaistettiin 22071 tervettä henkilöä aspiriiniryhmään ja lumeryhmään. Aspiriiniryhmään kuuluvat saivat päivittäin pienen annoksen aspiriinia. Henkilöiden terveydentilaa seurattiin keskimäärin 5 vuotta.

Asetelma on satunnaistettu kliininen koe, jossa tutkittiin aspiriinin käytön mahdollisesti alentavaa vaikutusta sydäninfarktikuolleisuuteen. Joka toinen päivä lääkärit ottivat aspiriinin tai lumetabletin. Tutkimukseen osallistuneet eivät tienneet, kumpaan ryhmään he kuuluivat.

Pääkysymys on siis tämä: Onko aspiriinista hyötyä sydäninfarktin ehkäisyssä? Aspiriiniryhmässä on vähemmän sydäninfarkteja kuin lumeryhmässä, 139 vastaan 239. Mitä tämä todistaa, mitä luvuista voidaan päätellä? Onko todistusaineisto kyllin vahva, jotta voimme vastata kysymykseen? Sivuvaikutuksia, jos niitä mitataan aivohalvausten määrällä, oli enemmän aspiriiniryhmässä. Kuitenkaan lukujen 119 vastaan 98 osoittama ero ei tunnu vakuuttavalta. Tällaiseen kysymykseen vastaaminen edellyttää *tilastollista* eli *stokastista mallia*, joka kuvaa havintojen stokastista käyttäytymistä.

Taulukko 7.1 Aspiriinin käytön mahdollinen aivohalvauksia ja sydäninfarkteja ehkäisevä vaikutus.

Ryhmä	Sydänkohtaus	Aivohalvaus	Yhteensä
Aspiriini	139	119	11037
Lume	239	98	11034
Yhteensä	378	217	22071

Suhteellinen riski

$$\frac{139/11037}{239/11034} = 0.58$$

on eräs vakiintunut vertailla kahta suhteellista osuutta. Aspiriinin hyöty suhteellisena riskinä on 0.58. Suhteellinen riski 1 tarkoittaa, että aspiriinista ei ole hyötyä. Ykköstä selvästi pienempi arvo osoittaa, että aspiriinista on hyötyä. Onko siis 0.58 tarpeeksi paljon ykköstä pienempi? Tässä esimerkissä voidaan olettaa binomimalli, jossa sydäninfarktiin sairastuneiden lukumäärä aspiriiniryhmässä noudattaa binomijakaumaa $\mathrm{Bin}(\theta_1, n_1)$ ja lumeryhmässä binomijakaumaa $\mathrm{Bin}(\theta_2, n_2)$, missä $n_1 = 11037, n_2 = 11034$ ja todennäköisyydet θ_1 sekä θ_2 ovat tuntemattomia parametreja. Tuntematon suhteellinen riski on $\theta_1/\theta_2 \equiv \theta$.

Olemme siis laskeneet suhteellisen riskin estimaatin $\hat{\theta} = 0.58$. Siihen ei ole liitetty mitään arvon luotettavuuta luonnehtivaa mittaa, joten se ei yksin pysty vastaamaan alkuperäiseen kysymykseen. Antaako koe niin paljon informaatiota, että voimme estimaatin $\hat{\theta}$ perusteella väittää θ:n olevan paljon pienempi kuin 1. Ajatellaan, että olisi tehty 10 kertaa laajempi koe ja olisi havaittu 1390 vastaan 2390 sydänkohtausta. Silloin jälleen olisi $\hat{\theta} = 0.58$, mutta intuitiivisesti tämän kokeen tulos tuntuu vakuuttavammalta. Estimaattiin pitää liittää jokin sen täsmällisyyttä kuvaava mitta, jonka avulla voimme arvioida estimaatin luotettavuutta. Tässä on päädytty tilastollisen päättelyn perusongelmaan: Miten havaintojen avulla voidaan tehdä meitä kiinnostavia tuntemattomia parametreja koskevia johtopäätöksiä?

Jos esimerkiksi todennäköisyys sairastua johonkin tautiin muuttuu jossain väestössä vaikkapa todennäköisyydestä 0.0001 todennäköisyyteen 0.0010, on muutos suhteellisesti ottaen erittäin suuri. Jonkin tavallisen tapahtuman kohdalla yhtä suuri todennäköisyyden muutos, esimerkiksi todennäköisyydestä 0.2001 todennäköisyyteen 0.2010 ei ole merkittävä. Samoilla todennäköisyyksien muutoksilla lähellä ääripäitä 0 ja 1 on usein suurempi merkitys kuin vaihtelualueen keskivaiheilla. Todennäköisyyksien suhteen avulla voidaan tarkastella suhteellista muutosta.

Taulukossa 7.2 *suhteellinen riski* on

$$\theta = \theta_1/\theta_2,$$

joka voi saada minkä tahansa ei-negatiivisen arvon. Todennäköisyyksien arvoilla $\theta_1 = 0.0010$ ja $\theta_2 = 0.0001$ suhteellinen riski on $SR = 0.0010/0.0001 = 10$ ja arvoilla $\theta_1 = 0.2010$ ja $\theta_2 = 0.2001$ suhteellinen riski on $SR = 0.2010/0.2001 = 1.004$.

Taulukko 7.2 Sairastumistodennäköisyys Aspiriiniyhmässä on θ_1 ja Lumeryhmässä θ_2. Indikaattori $Y = 1$, kun henkilöllä infarkti ja muutoin $Y = 0$, joten $P(Y = 1|\text{Aspiriiniryhmä}) = \theta_1$ ja $P(Y = 1|\text{Lumeryhmä}) = \theta_2$.

	Y		
	1	0	Yhteensä
Aspiriini	θ_1	$1 - \theta_1$	1.0
Lume	θ_2	$1 - \theta_2$	1.0

7.2 Tilastolliset mallit

Tilastotiede käsittelee sitä, mitä voimme oppia havainnoista. Tilastollisen mallintamisessa tarkastellaan havaintoja satunnaiskokeen tuloksena, satunnaisilmiönä. Ajatelkaamme, että havainnot ovat "mustan laatikon"tuottamia. Syötemuuttujien vektori $\boldsymbol{x} = (x_1, \ldots, x_p)$ (selittävät muuttujat, riippumattomat muuttuja) menee sisään laatikkoon ja vastemuuttujat (riippuvat muuttujat, selitettävät muuttujat) $\boldsymbol{y} = (y_1, \ldots, y_m)$ tulevat ulos:

$$\boldsymbol{x} \longrightarrow \boxed{\text{luonto}} \longrightarrow \boldsymbol{y}$$

Laatikon sisällä luonto liittää jonkin funktion avulla selittävät muuttujat ja vasteet yhteen. Havaintojen analysoinnin tavoitteet voidaan karkeati jakaa kahteen ryhmään:

> *Ennustaminen.* Mallilla halutaan ennustaa, mitä vatemuuttujan arvoja saadaa tulevaisuuden syötteillä.
>
> *Tieto riippuvuuksista.* Halutaan saada selvyys siitä, miten luonto on liittänyt yhteen syötteet ja vastemuuttujat.

Tilastolliset tulokset ja näkemykset voidaan tasmällisimmin ilmaista matematiikan keinoin, mutta yhteys havaintoihin ja tieteelliseen päättelyyn on ominaista tilastotieteelliselle ajattelutavalle. Monet tilastolliset tulokset ovat syntyneet ja syntyvät vastauksena melko konkreettisiin kysymyksiin, joihin ei aina ole olemassa yleistä eleganttia ratkaisua. Tällaisten ongelmien valtava määrä tekee vaikeaksi kehittää kaiken kattavaa teoriaa, vaikka toisaalta yhteisiä yleisiä periaatteita voidaan esittää. Tällaiset periaatteet voidaan kiteyttää tilastollisen mallin käsitteeseen.

Havaintojen parametrinen mallintaminen

Tässä lahestymistavassa oletetaan mustan laatikon sisälle jokin havaintoja gereroiva todennäköisyysmalli. Havaintojen malli on esimerkiksi muotoa
(7.2.1)

$$\text{vastemuuttujat} = f(\text{selittävät muuttujat, satunnaisvirhe, parametrit}).$$

Parametrit estimoidaan havainnoista ja sen jälkeen mallia voidaan käyttää riippuvuuksien tarkasteluun tai ennustamiseen. Musta laatikko voisi näyttää esimerkiksi seuraavalta:

$$\boldsymbol{x} \longrightarrow \boxed{\text{lineaarinen regressio}} \longrightarrow y$$

Tässä siis kuvitellaan, että lineaarinen regressiomalli kuvaa riittävän hyvin vasteen y riippuvuutta selittäjistä $\boldsymbol{x}$. Malli pyritään vahvistamaan havaintojen avulla. Käytetään esimerkiksi yhteensopivuustestejä ja residuaalien tarkastelua.

Algoritminen mallintaminen

Algoritminen lähestymistapa on saavuttanut lisaa suosiota ja sovellusmahdollisuuksia tietokoneiden laskentakapasiteetin kasvun myötä. Tässä ajattelutavassa mustan laatikon sisältö on monimutkainen ja tuntematon. Pyritään löytämään funktio $f(\boldsymbol{x})$ algoritmisesti– algoritmi laskee $\boldsymbol{x}$:n perusteella ennusteita $\boldsymbol{y}$:lle ja pyrkii muokkaamaan algoritmin sellaiseksi, että se antaa hyviä ennusteita. Musta laatikko näyttäisi tältä:

$$\boldsymbol{x} \longrightarrow \boxed{\text{tuntematon}} \longrightarrow y$$

Esimerkiksi neuroverkot kuuluvat tähän kategoriaan. Mallin pätevyyttä arvioidaan ennustevirheen avulla.

Tavallisimmissa malleissa havaintojen vaihtelun jako

$$\begin{aligned}\text{havainnot} &= f(\text{selittävät muuttujat, parametrit}) + \text{satunnaisosa}\\ &= \text{systemaattinen osa} + \text{satunnaisosa},\end{aligned}$$

systemaattiseen osaan ja satunnaisosaan on additiivinen. Esitys (7.2.1) on itse asiassa varsin yleinen, joka sallii monimutkaisetkin vaikutusmekanismit. Yksinkertaistavia oletuksia kuitenkin tarvitaan, jotta mallit pystytään ymmätämään ja analysoimaan. Havaintojen oletetaan olevan peräisin jostain jakaumaperheestä, tavallisimmin ns. parametrisesta jakaumaperheestä. Systemaattinen osa on esimerkiksi havaintojen $Y_1, Y_2, \ldots, Y_n$ odotusarvoja $E(Y_i)$, $1 \leq i \leq n$, koskeva oletus, joka lausutaan vaikkapa regressiofunktiona. Tavallisesti odotusarvo riippuu joistain selittävistä muuttujista (eli kovariaatista). Tilastollisen mallin voidaan sanoa olevan havaintojen yhteisjakaumaa ja systemaattista osaa koskevien oletusten joukko.

Esimerkki 7.1 Oletetaan, että $Y_1, Y_2, \ldots, Y_n$ on otos normaalijakaumasta $\mathrm{N}(\mu_i, \sigma^2)$, missä $E(Y_i) = \mu_i$. Oletetaan lisäksi, että $E(Y_i)$ riippuu lineaarisesti selittävästä muuttujasta x, joten

$$\mu_i = \alpha + \beta x_i, \qquad 1 \leq i \leq n. \tag{7.2.2}$$

Malli voidaan kirjoittaa myös muodossa

$$Y_i = \mu_i + \varepsilon_i,$$

missä $\varepsilon_i = Y_i - E(Y_i)$. Virhetermi ε_i noudattaa normaalijakaumaa $\mathrm{N}(0, \sigma^2)$.

Kokeellisessa tilanteessa selittäjä x on koevakio. Tutkija voi päättää, mitkä arvot hän x:lle valitsee. Esimerkiksi törmäystestissä valitaan törmäysnopeudet $x_1, \ldots, x_n$. Näillä selittäjän arvoilla mitataan vastemuuttujan (tai vastemuuttujien) arvot. Regressioanalyysia käytetään kuitenkin myös ei-kokeellisessa tilanteessa, jossa tutkija ei voi kontrolloida x:n arvoja. Silloin x on satunnaismuuttuja, jonka arvo havainnoidaan usein samanaikaisesti vastemuuttujan kanssa. On huomattava, että regressiomallissa (7.2.2) tarkastellaan ehdollista odotusarvoa

$$E(Y|x) = \mu(x),$$

missä Y:n ehdollisen odotusarvon oletetaan olevan x:n linearinen funktio $\mu(x) = \alpha + \beta x$. Suoran kertoimet α ja β ovat tuntemattomia parametreja, jotka estimoidaan havainnoista. □

Tilastotieteen oppikirjoissa lähdetään tavallisesti liikkeelle melko teknisesti. Sanotaan, että havainnot $Y_1, \ldots, Y_n$ ovat otos jostain tuntenmattomasta jakaumasta F, missä F on siis jakauman kertymäfunktio. Tavallisesti jakaumasta tehdään joitain oletuksia. Tilanne voi olla esimerkiksi sellainen, että jakauma voidaan oletettaa symmetriseksi. Tällä kurssilla käytetään useimmiten parametrista lähestymistapaa, jolloin jakauman ajatellaan kuuluvan johonkin parametriseen funktioperheeseen

$$\mathcal{F} = \{\, F(x;\theta),\ \theta \in \Theta \,\}$$

missä $F(x;\theta)$ on kertymäfunktio jokaisella kiinnitetyllä θ:n arvolla. Käsittelemissämme päättelyongelmissa operoimme tavallisesti tiheysfunktioiden avulla, joten jakaumaperhe on silloin suoraviivaisempaa luonnehtia tiheysfunktioiden joukkona

$$\mathcal{F} = \{\, f(x;\theta),\ \theta \in \Theta \,\}.$$

Suure θ on siis parametri ja sen arvojoukko Θ on parametriavaruus. Valitsemalla yksi parametrin θ arvo saadaan täysin määrätty jakauma. Edellä olemme nähneet, että θ voi olla selittäjien funktio. Kun parametrin θ arvo valitaan havaintojen perusteella, saadaan θ:n *piste-estimaatti*. Parametrin (parametrien) arvo määrittämistä havaintojen perusteella sanotaan *piste-estimaattoinniksi*.

Taulukko 7.1. Vakavien onnettomuuksien lukumäärä alueella A tammikuussa vuonna 2000.

Yli 21-vuotiaat		Alle 21-vuotiaat	
Kuolemaan johtaneet	Muut	Kuolemaan johtaneet	Muut
Y_1	Y_2	Y_3	Y_4
11	62	4	7

Esimerkki 7.2 Tarkastellaan auto-onnettomuuksien vakavuusastetta, kun selittäjänä on kuljettajan ikä. Usein väitetään, että nuoret kuljettajat aiheuttavat keskimääräistä enemmän vakavia onnettomuuksia.

Oletetaan, että onnettomuuksien lukumäärä kuukaudessa noudattaa Poissonin jakaumaa Poi(λ). Tarkastellaan neljää onnettomuustyyppiä, jotka on määritelty kuljettajan iän ja onnettomuuden vakavuusasteen mukaan. Eri tyyppisten onnettomuuksien lukumäärien $Y_i, 1 \leq i \leq 4$, oletetaan noudattavan toisistaan riippumatta Poissonin jakaumaa Poi(λ_i). Oheisessa taulukossa on annettu eräs aineisto. Silloin esimerkiksi kuolemaan johtaneiden onnettomuuksien lukumäärä Y_3 alle 21-vuotiaiden ryhmässä noudattaa Poissonin jakaumaa Poi(λ_3). Parametrit λ_1, λ_2, λ_3 ja λ_4 ovat satunnaismuuttujien Y_1, Y_2, Y_3 ja Y_4 odotusarvoja. Odotusarvo λ_i kertoo onnettomuusasteen i. kategoriassa. Vastaavasti esimerkiksi yli 21-vuotiaiden onnettomuusaste on $\lambda_1 + \lambda_2$ ja alle 21-vuotiaiden $\lambda_3 + \lambda_4$. Merkitään $\theta_1 = \lambda_1 + \lambda_2$ ja $\theta_2 = \lambda_3 + \lambda_4$. Näin todennäköisyys, että yli 21-vuotias aiheuttaa kohtalokkaan onnettomuuden, on

$$\pi_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

ja alle 21-vuotiaan todennäköisyys aiheuttaa kohtalokas onnettomuus on

$$\pi_2 = \frac{\lambda_3}{\lambda_3 + \lambda_4}.$$

Nelikko $(\theta_1, \pi_1, \theta_2, \pi_2)$ muodostaa uuden parametrisoinnin, joka saattaa olla tulkinnallisesti selkeämpi ja mielenkiintoisempi kuin alkuperäinen. □

7.3 Estimoinnista

Tarkastelemme tässä luvussa satunnaismuuttujia, joiden todennäköisyysfunktion (tai tiheysfunktion) funktionaalinen muoto tunnetaan, mutta jakauma riippuu jostain tuntemattomasta parametristä θ. Oletetaan, että parametrin θ mahdolliset arvot kuuluvat johonkin annettuun joukkoon Θ, jota kutsutaan *parametriavaruudeksi*. Tiedetään esimerkiksi, että jonkin tuotteen elinaika X noudattaa eksponenttijakaumaa

$$f(x;\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}, \qquad 0 < x < \infty,$$

missä $\theta \in \Theta = \{\, \theta \mid 0 < \theta < \infty \,\}$. Parametriavaruus Θ on siis positiivisten reaalilukujen joukko. Haluamme valita funktioperheestä

$$\mathcal{F} = \{\, f(x;\theta),\ \theta \in \Theta \,\}$$

yhden tiheysfunktion, joka esittää parhaiten tuotteen elinaikaa. Valitaan siis yksi parametrin θ arvo eli parametrin θ *piste-estimaatti*, joka määrittää jakauman.

Parametrin arvo arvioidaan eli estimoidaan havaintojen perusteella. Teemme jakaumasta havainnon $X = x$ ja estimoimme parametrin θ arvon havainnon x perusteella. Parametrin θ estimoimiseen käytettävää otosfunktiota $T(X)$ kutsutaan parametrin θ *estimaattoriksi* ja estimaattorin $T(X)$ arvoa $t = T(x)$ kutsutaan parametrin θ *estimaatiksi*. Estimaattori pyritään valitsemaan siten, että se antaa hyviä arvioita parametrista θ.

Esimerkki 7.3 Estimoidaan ehdokkaan A kannattajien suhteellinen osuus π eräässä suuressa kaupungissa. Valitaan kaupungin äänioikeutetuista satunnaisesti n henkilöä, joilta tiedustellaan, kannattavatko he ehdokasta A. Olkoon X ehdokkaan A kannattajien lukumäärä otoksessa. Koska populaation koko on suuri verrattuna otoskokoon n, voidaan olettaa, että $X \sim \mathrm{Bin}(n, \pi)$, missä π on todennäköisyys, että satunnaisesti valittu henkilö kannattaa A:ta. Binomijakaumaa noudattavan satunnaismuuttujan X todennäköisyysfunktio on muotoa

$$f(x; n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \qquad x = 0, 1, \ldots, n; \quad 0 \le \pi \le 1.$$

Binomijakauman parametriavaruus on $\Theta = \{\, \pi \mid 0 \le \pi \le 1 \,\}$. Tehtävänämme on määrittää π:n estimaattori $T(X)$ siten, että havaitun arvon $X = x$ perusteella saadaan hyvä π:n piste-estimaatti $T(x)$. Havainnon $X = x$ todennäköisyys on

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}. \tag{7.3.1}$$

Eräs tapa määrittää π:n estimaatti on tarkastella todennäköisyyttä $P(X = x)$ parametrin π funktiona ja etsiä sellainen π:n arvo, että havainnon x todennäköisyys saavuttaa maksiminsa. Voidaan osoittaa, että havainnon $X = x$ todennäköisyys maksimoituu, kun $\pi = x/n$. Tätä estimaattia kutsutaan π:n *suurimman uskottavuuden estimaatiksi* ja sitä merkitään

$$\hat{\pi} = \frac{x}{n}.$$

□