Luku 1

Johdanto

Tilastollinen päättely (engl. *statistical inference*) on kokoelma käsitteitä ja menetelmiä. Niiden tarkoitus on auttaa soveltajaa tekemään päätelmiä reaalimaailman olosuhteista, kun näitä olosuhteita ei havaita suoraan, vaan päätelmät pitää tehdä epävarmuutta sisältävien numeeristen havaintojen perusteella.

Matemattinen päättely on luonteeltaan *deduktiivista*: yleisistä säännöistä (aksioomeista) päätellään niiden seurauksia. Tästä poiketen tilastollinen päättely on luonteeltaan *induktiivista*: siinä pyritään yksittäisistä havainnoista kohti yleisiä sääntöjä. Tilastollinen päättely on luonteeltaan epävarmaa ja sisältää aina virheellisen päättelyn mahdollisuuden. Tämän epävarmuuden suuruutta on kuitenkin mahdollista kontrolloida ja arvioida.

Tilastollisessa päättelyssä käytetään hyväksi matematiikkaa, erityisesti todennäköisyyslaskentaa, mutta tilastollinen päättely ei ole matematiikan vaan tilastotieteen osa-alue. Tilastollinen päättely on todennäköisyyslaskennalle *käänteinen ongelma*: todennäköisyyslaskenta tarjoaa työkaluja, joilla voidaan laskea havaintojen jakauma tai niistä laskettujen tilastollisten tunnuslukujen jakauma, kun havaintoja generoiva todennäköisyysmalli on kiinnitetty. Tilastollisessa päättelyssä pitää numeerisen aineiston perusteella yrittää arvioida, minkälainen todennäköisyysmalli olisi ne voinut generoida.

Tilastotieteen soveltajat elävät usein sellaisessa harhaluulossa, että tilastollinen päättelyn oppikirjat ovat keittokirjoja, joista löytyy sopiva resepti (menetelmä) kunkin empiirisen tieteen tutkimusongelman ratkaisemista varten. Tämä ei pidä paikkaansa. Alan oppikirjoista toki löytyy tiettyjä usein sovelluksissa käytettäviä reseptejä (menetelmiä), mutta ne perustuvat aina tiettyihin oletuksiin. Kussakin tilastollisen menetelmän sovelluksessa pitää erikseen kriittisesti arvioida, toteutuvatko kyseisen menetelmän oletukset. Mikäli oletukset eivät täyty, saattaa tilanteeseen sopivan menetelmän rakentelu vaatia pitkän tutkimushankkeen. Sitä paitsi tilastollisen päättelyn ideaa voidaan lähestyä kahdesta aivan erilaisesta lähtökohdasta, joista keittokirjamaisissa oppikirjoissa tavallisesti esitetään vain toinen.

Tilastolliseen päättelyyn on olemassa kaksi periaatteiltaan erilaista lähestymistapaa: frekventistinen päättely sekä bayesiläinen päättely. Tällä kurssilla käsitellään enimmäkseen frekventististä päättelyä. Sen avulla saadaan tietyissä yksinkertaisissa tilanteissa helposti sovellettavia menetelmiä, jotka ovat laajalti tunnettuja.

Tarkempi tarkastelu paljastaa kuitenkin, että tietyt frekventistisen lähesty-

mistavan periaatteet ovat ongelmallisia, ja tämä voi johtaa käytännön ongelmiin monimutkaisissa tilanteissa. Bayesiläinen lähestymistapa perustuu puhtaasti todennäköisyyslaskennan soveltamiseen, ja se on tämän matemaattisen muotoilun ansiosta matemaattisesti selkeää sekä vapaa tietyistä frekventististä lähestymistapaa vaivaavista käsitteellisistä ongelmista. Vaikka bayesiläisen päättelyn matemaattinen muotoilu on selkeää, niin sen sijaan siinä sovellettava todennäköisyyskäsitteen tulkinta kvantitatiivisenä esityksenä tutkijan epävarmuudesta on joidenkin mielestä ongelmallinen. Valitettavasti baysiläinen päättely vaatii hieman laajempia tietoja todennäköisyyslaskennasta kuin mitä tämän kurssin opiskelijoilta oletetaan, minkä takia bayesiläistä päättelyä käsitellään tällä kurssilla vain ylimalkaisesti.

Tilastollisen päättelyn oppikirjoja on olemassa satoja ellei tuhansia. Tässä monisteessa ei pyritä esittämään mitään omintakeista, vaan tässä käydään läpi alan peruskäsitteitä, minkä takia en esitä yksityiskohtaisia kirjallisuusviitteitä. Näitä luentomuistiinpanoja laatiessani olen ottanut eniten mallia (ts. varastanut sumeilematta materiaalia) tätä kurssin versiota edeltävän kurssin version luentomuistiinpanoista, jotka laati E. Arjas yhdessä J. Sirénin kanssa. Lisäksi olen tarkistanut, kuinka T. Mäkeläinen aikanaan esitti vastaavat asiat omassa luentomonisteessaan. Tämän lisäksi olen katsonut, kuinka P. Nieminen ja P. Saikkonen esittävät tilastollisen päättelyn perusteet Tilastollisen päättelyn kurssin kurssimonisteessa. Olen myös pitänyt käsillä mm. seuraavia englanninkielisiä oppikirjoja: Arnold [1], Casella ja Berger [2], Davison [3], Ross [4]. Todennäköisyyslaskennan osalta oletan lukijalla olevan suunnilleen Pekka Tuomisen kirjaa [5] vastaavat tiedot.

Nykyaikana tilastollisen päättelyn vaatimat laskut toteutetaan tietokoneella. Joissakin kohdissa olen näyttänyt, kuinka laskut saataisiin toteutettua R-tilastoohjelmistossa.

Kirjallisuutta

[1] S. F. Arnold. *Mathematical Statistics.* Prentice-Hall, Inc., 1990.

[2] George Casella and Roger L. Berger. *Statistical Inference.* Duxbury Press, 2nd edition, 2002.

[3] A. C. Davison. *Statistical Models.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2003.

[4] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists.* Elsevier Academic Press, 4th edition, 2009.

[5] P. Tuominen. *Todennäköisyyslaskenta I.* Limes ry, Helsinki, 1993.

Luku 2

Havaintojen mallintaminen

2.1 Havaintoja vastaava todennäköisyysmalli

Meillä on käsillä numeerinen *aineisto* (engl. *data*) $y_1, \ldots, y_n$, jossa kukin y_i on jokin tunnettu luku. Havaintojen lukumäärää n kutsutaan *otoskooksi* (engl. *sample size*). Ennen havaintojen tekoa aineiston arvot ovat epävarmoja (mittausvirheiden, koetilanteessa tehdyn satunnaistamisen, populaation luonnollisen vaihtelun tms. syyn takia). Kokeen tai otannan toistaminen voisi tuottaa toisenlaiset havainnot. Tämän takia mallinnamme tilanteen niin, että arvot $y_1, \ldots, y_n$ ovat satunnaismuuttujien $Y_1, \ldots, Y_n$ toteutuneita arvoja (eli niiden reaalisaatioita).

> Tämä on tilastollisen päättelyn perusajatus: havaittujen arvojen ajatellaan olevan satunnaismuuttujien toteutuneita arvoja.

Satunnaismuuttujat ovat jollakin perusjoukolla Ω määriteltyjä reaaliarvoisia funktioita, joten edellisen mukaan ajattelemme, että

$$y_1 = Y_1(\omega^{\text{act}}), y_2 = Y_2(\omega^{\text{act}}), \ldots, y_n = Y_n(\omega^{\text{act}}), \tag{2.1}$$

jossa $\omega^{\text{act}} \in \Omega$ on todennäköisyysmallissa aktualisoitunut alkeistapaus, jonka luontoäiti (tms. epämääräiseksi jäävä taho) on valinnut.

Otamme merkintöjen lyhentämiseksi käyttöön vektorimerkinnät sekä aineistolle että aineistoa vastaaville satunnaismuuttujille,

$$\mathbf{y} = (y_1, \ldots, y_n), \qquad \mathbf{Y} = (Y_1, \ldots, Y_n),$$

Tässä $\mathbf{y} \in \mathbb{R}^n$ on havaituista arvoista muodostettu havaintovektori tai aineisto, ja $\mathbf{Y}$ on havaintovektoria $\mathbf{y}$ vastaava satunnaisvektori, eli havaintosatunnaisvektori. Matemaattisesti $\mathbf{Y}$ on kuvaus $\Omega \to \mathbb{R}^n$, ja mallimme mukaan

$$\mathbf{y} = \mathbf{Y}(\omega^{\text{act}})$$

jollekin $\omega^{\text{act}} \in \Omega$.

> Tilastollisen päättelyn tavoitteena on tehdä aineiston $\mathbf{y}$ perusteella johtopäätöksiä siitä todennäköisyysjakaumasta, jota satunnaisvektori $\mathbf{Y}$ noudattaa.

Tyypillisesti vektorin $\mathbf{Y}$ jakauma mallinnetaan parametrisella mallilla, jossa on yksi parametri θ, tai monimutkaisemmissa tilanteissa useampia parametreja $\theta_1, \ldots, \theta_p$, joista yhdessä muodostuu parametrivektori $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Tällä kurssilla oletamme, että kaikki satunnaismuuttujat Y_i ovat joko diskreettejä (jolloin niistä kunkin jakaumaa kuvaa pistetodennäköisyysfunktio) tai että kaikki satunnaismuuttujat Y_i ovat jatkuvasti jakautuneita (jolloin niistä kunkin jakaumaa kuvaa tiheysfunktio). Kun parametrin (tai yleisemmässä tapauksessa parametrivektorin) arvo on kiinnitetty, niin satunnaisvektorin $\mathbf{Y}$ jakauman esittää sen yhteispistetodennäköisyysfunktio (yptnf) tai yhteistiheysfunktio (ytf)

$$f(\mathbf{y}; \theta) = f(y_1, \ldots, y_n; \theta)$$

Tämä yptnf/ytf riippuu $n+1$ reaalimuuttujasta $y_1, \ldots, y_n, \theta$, joista θ on merkitty puolipisteen jälkeen, koska se on erilaisessa roolissa kuin muuttujat $y_1, \ldots, y_n$. Muuttuja $\mathbf{y} = (y_1, \ldots, y_n)$ on vapaa muuttuja, eikä tässä kaavassa eikä monessa muussakaan kaavassa vielä tarkoita aineistoa. Saman symbolin käyttäminen selkeästi eri merkityksissä on tilastotieteen merkinnöille tyypillistä, ja siihen on lukijan parasta vain totuttautua. Kullakin kiinteällä θ funktio

$$\mathbf{y} \mapsto f(\mathbf{y}; \theta)$$

on yptnf tai ytf

Tällä kurssilla käytetään lähes yksinomaan sellaisia malleja, joissa satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat riippumattomia, kun parametrin arvo on kiinnitetty. Tällaisessa tilanteessa yptnf/ytf voidaan esittää tulona kaavalla

$$f(\mathbf{y}; \theta) = f_{Y_1}(y_1; \theta)\, f_{Y_2}(y_2; \theta) \cdots f_{Y_n}(y_n; \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta) \tag{2.2}$$

jossa $f_{Y_i}(u; \theta)$ tarkoittaa satunnaismuuttujan Y_i pistetodennäköisyysfunktiota (ptnf) tai tiheysfunktiota (tf), kun parametrilla on arvo θ.

Usein käsittelemme tilannetta, jossa satunnaismuuttujat Y_i ovat riippumattomia ja niillä on sama jakauma, kun parametrinarvo θ on kiinnitetty. Tässä tapauksessa sanotaan, että satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat *satunnaisotos* (engl. *random sample*) ko. jakaumasta. Jos tämän yhteisen jakauman tiheysfunktio (tf) tai pistetodennäköisyysfunktio (ptnf) on $g(y; \theta)$, niin kaavasta (2.2) saadaan yhteisjakaumalle esitys

$$f(\mathbf{y}; \theta) = g(y_1; \theta) \cdots g(y_n; \theta) = \prod_{i=1}^{n} g(y_i; \theta). \tag{2.3}$$

Tilastollisessa päättelyssä kiinnostuksen kohteena on sv:n $\mathbf{Y}$ jakauma, ja parametrisessa mallissa kyseinen jakauma tunnetaan täysin, jos parametrinarvo θ tunnetaan. Ongelma syntyy siitä, että θ on tuntematon. Parametrista tiedetään kuitenkin vähintään sen verran, että osataan sanoa, missä joukossa $\Theta \subset \mathbb{R}$ sen arvot voivat olla. Tällaista joukkoa Θ kutsutaan *parametriavaruudeksi* (engl. *parameter space*).

Tällä kurssilla havaintoja kuvaavaa todennäköisyysmallin $f(\mathbf{y}; \theta)$ ajatellaan enimmäkseen olevan valmiiksi annettu. Käytännössä sovelletaan usein konventionaalisia malleja, joiden ominaisuudet tunnetaan hyvin.

Mallin pitäisi toki vastata todellisuutta. Malleissa yleensä oletetaan, että jotkin niissä esiintyvät satunnaismuuttujat ovat riippumattomuutta. Tällaista riippumattomuusoletusta on mahdotonta tarkistaa numeerisesta aineistosta: luvut eivät ole toisistaan riippumattomia, vaan riippumattomuus on satunnaismuuttujien ominaisuus. Riippumattomuusoletuksia pitäisi pohtia kriittisesti käyttämällä hyväksi sitä tietoa, mikä on käytössä koeasetelmasta. Mikäli mahdollista, koeasetelma pitäisi suunnitella etukäteen niin, että se mahdollisimman hyvin toteuttaa päättelyssä käytettävän mallin oletukset.

Yleisesti ottaen havaintojen mallintaminen on vaativa tehtävä. Tarkastelemme kuitenkin seuraavaksi kahta esimerkkiä, joissa todennäköisyysmallin $f(\mathbf{y};\theta)$ muodostaminen on lähes itsestään selvää.

2.2 Pallot kulhossa

Oletamme, että kulhossa on samankokoisia ja samasta materiaalista valmistettuja valkoisia ja mustia palloja yhteensä N kappaletta. Merkitään valkoisten pallojen lukumäärää $\theta = \#\{\text{valkoiset pallot}\}$, jolloin kulhossa on $N - \theta$ mustaa palloa. Oletamme, että N on tunnettu luku, mutta θ on tuntematon. Parametriavaruus on $\{0, 1, \ldots, N\}$.

Kulhoa ravistetaan tarmokkaasti, ja sitten siitä nostetaan yksi pallo sokkona. Koska kulhossa on yhteensä N palloa, ja niistä θ on valkoista, niin on luonnollista ajatella, että

$$P_\theta(\text{nostettu pallo on valkoinen}) = \frac{\theta}{N}.$$

Edellä merkittiin parametri θ selvyyden vuoksi näkyviin todennäköisyyden $P(\cdot)$ alaindeksiksi. Jotta edellä kirjoitetulla todennäköisyydellä olisi numeerinen arvo, täytyy luvun N sekä valkoisten pallojen lukumäärän θ olla tunnettuja lukuja.

Tarkastelemme seuraavaksi poimintaa takaisinpanolla (eli palauttaen). Nostettu pallo palautetaan kulhoon, kulhoa ravistetaan ja nostetaan toinen pallo sokkona. Tätä menettelyä toistetaan n kertaa, niin että nostettu pallo aina palautetaan kulhoon noston jälkeen ja ennen kutakin nostoa kulhoa ravistetaan perusteellisesti.

Määrittelemme satunnaismuuttujan Y_i kullekin $i = 1, \ldots, n$ seuraavalla tavalla:

$$Y_i = \begin{cases} 1, & \text{jos } i\text{:nnellä nostolla saadaan valkoinen pallo,} \\ 0, & \text{jos } i\text{:nnellä nostolla saaadaan musta pallo.} \end{cases}$$

Voimme kirjoittaa

$$\begin{aligned} P_\theta(Y_i = 1) &= \theta/N \\ P_\theta(Y_i = 0) &= 1 - \theta/N. \end{aligned}$$

Nämä tulokset voidaan esittää myös yhdellä kaavalla

$$P_\theta(Y_i = y_i) = \left(\frac{\theta}{N}\right)^{y_i} \left(1 - \frac{\theta}{N}\right)^{1-y_i}, \qquad y_i = 0, 1.$$

Tämä lauseke on satunnaismuuttujan Y_i pistetodennäköisyysfunktio $f_{Y_i}(y_i;\theta)$.

Koska kulhoa aina ravistetaan perusteellisesti ennen kutakin nostoa ja koska nostetut pallot aina palautetaan kulhoon, niin on luonnollista ajatella, että nostoja vastaavat satunnaismuuttujat ovat riippumattomia, koska arkijärjen mukaan tieto yhden noston lopputuloksesta ei voi vaikuttaa toisen noston todennäköisyysjakaumaan. Satunnaismuuttujien yptnf on kaavan (2.2) tai sen erikoistapauksen (2.3) mukaisesti

$$\begin{aligned} f(\mathbf{y};\theta) &= f(y_1,\ldots,y_n;\theta) \\ &= f_{Y_1}(y_1;\theta)\, f_{Y_2}(y_2;\theta)\cdots f_{Y_n}(y_n;\theta) \\ &= \left(\frac{\theta}{N}\right)^{y_1}\left(1-\frac{\theta}{N}\right)^{1-y_1}\left(\frac{\theta}{N}\right)^{y_2}\left(1-\frac{\theta}{N}\right)^{1-y_2}\cdots\left(\frac{\theta}{N}\right)^{y_n}\left(1-\frac{\theta}{N}\right)^{1-y_n} \end{aligned}$$

Kukin y_i saa joko arvon 0 tai 1 ja parametrin θ arvo on jokin luvuista $0, 1, \ldots, N$. Jatkokehittelyjä varten on hyödyllistä huomata, että yptnf voidaan esittää (yhdistämällä termien θ ja $(1-\theta)$ potenssit) myös muodossa

$$f(\mathbf{y};\theta) = \left(\frac{\theta}{N}\right)^{t(\mathbf{y})}\left(1-\frac{\theta}{N}\right)^{n-t(\mathbf{y})}, \tag{2.4}$$

jossa $t(\mathbf{y}) = y_1 + \cdots + y_n$ on yhteensä n nostolla saatu valkoisten pallojen lukumäärä (onnistumisten lukumäärä) ja $n - t(\mathbf{y})$ on yhteensä n nostolla saatu mustien pallojen lukumäärä (epäonnistumisten lukumäärä).

Tässä (ja kaikissa muissakin esimerkeissä) olisi yhteisjakauma voitu parametroida myös toisela tavalla. Esimerkiksi parametriksi voitaisiin ottaa valkoisten pallojen suhteellinen osuus kulhossa olevista palloista. Jos θ on valkoisten pallojen lukumäärä kulhossa, niin niiden suhteellinen osuus on

$$\phi = \theta/N,$$

ja tämän parametrin avulla esitettynä aineistoa vastaavan satunnaisvektorin jakauman esittää yptnf

$$f_1(\mathbf{y};\phi) = f(\mathbf{y};\theta/N) = \phi^{t(\mathbf{y})}\,(1-\phi)^{n-t(\mathbf{y})}.$$

Uutta parametrointia vastaava parametriavaruus on joukko

$$\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}.$$

Kumpikin parametrointi on yhtä lailla oikea. Se mitä parametrointia kussakin tehtävässä käytetään on makuasia.

Tässä esimerkissä parametrin θ todellinen arvo voitaisiin selvittää katsomalla kulhoon. Kokeen lopputuloksen perusteella saattaa olla mahdollista sulkea pois tiettyjä parametrinarvoja. Mikäli yhdessäkin nostossa saadaan valkoinen pallo, niin arvo $\theta = 0$ voidaan sulkea pois. Vastaavasti, jos yhdessäkin nostossa saadaan musta pallo, niin arvo $\theta = N$ voidaan sulkea pois. Kuvatun koejärjestelyn puitteissa parametrin todellista arvoa ei kuitenkaan voida selvittää täysin varmasti oli nostojen lukumäärä n miten suuri hyvänsä (mikäli $N \geq 3$).

Pallojen palauttaminen kulhoon on välttämätöntä, jotta nostojen tuloksia voitaisiin pitää riippumattomina. Jos ensimmäistä palloa ei palauteta kulhoon,

niin kahden ensimmäisen noston tuloksille saamme mallin

$$\begin{aligned} P_\theta(Y_1 = 1, Y_2 = 1) &= P_\theta(Y_1 = 1)\, P_\theta(Y_2 = 1 \mid Y_1 = 1) \\ &= \left(\frac{\theta}{N}\right)^{y_1} \left(1 - \frac{\theta}{N}\right)^{1-y_1} \left(\frac{\theta - 1}{N - 1}\right)^{y_2} \left(1 - \frac{\theta - 1}{N - 1}\right)^{1-y_2} \end{aligned}$$

sillä jos ensin nostetaan valkoinen pallo, niin sen jälkeen kulhossa on jäljellä $N - 1$ palloa, joista $\theta - 1$ on valkoista. Jos taas ensin nostetaan musta pallo, niin tällöin

$$\begin{aligned} P_\theta(Y_1 = 0, Y_2 = 1) &= P_\theta(Y_1 = 0)\, P_\theta(Y_2 = 1 \mid Y_1 = 0) \\ &= \left(\frac{\theta}{N}\right)^{y_1} \left(1 - \frac{\theta}{N}\right)^{1-y_1} \left(\frac{\theta}{N - 1}\right)^{y_2} \left(1 - \frac{\theta}{N - 1}\right)^{1-y_2} \end{aligned}$$

Poiminnassa ilman takaisinpanoa aikaisemman noston lopputulos vaikuttaa seuraavan noston todennäköisyysjakaumaan, joten nyt Y_1 ja Y_2 eivät ole enää riippumattomia (kun θ:n arvo on kiinnitetty). Samaa järkeilyä voitaisiin jatkaa useammalle kuin kahdelle nostolle.

2.3 Nasta purkissa

Purkissa on nasta. Purkkia ravistetaan tarmokkaasti, ja sitten merkitään muistiin, laskeutuuko nasta selälleen vai kyljelleen. Tätä koetta toistetaan n kertaa.

Otamme käyttöön satunnaismuuttujat Y_i siten, että

$$Y_i = \begin{cases} 1, & \text{jos } i\text{:nnessä toistossa nasta päätyy selälleen,} \\ 0, & \text{jos } i\text{:nnessä toistossa nasta päätyy kyljelleen.} \end{cases}$$

Tuntuu luontevalta ajatella, että parametriksi valitaan välillä $(0, 1)$ oleva luku θ, joka tulkitaan todennäköisyydeksi, jolla nasta päätyy yhdessä toistossa selälleen. Tätä parametria ei voida selvittää purkkia ja nastaa katsomalla. Voidaan ajatella, että θ olisi yhtä kuin selälleen päätyvien tulosten suhteellinen osuus äärettömän pitkässä koesarjassa. Millään äärellisen pitkällä koesarjalla θ:n arvoa ei saada täydellisesti selville.

Tätä mallia voidaan kritisoida. On aivan ilmeistä, että ravistustapa vaikuttaa oleellisella tavalla lopputulokseen. Jos purkkia ravistetaan vain hitusen, niin nastan tila ei vaihdu. Tämän takia vaadimme, että ravistus on niin tarmokas, että nasta poukkoilee purkissa monta kertaa ympäriinsä seinästä toiseen. Se kumpi lopputulos kulloinkin saadaan olisi periaatteessa laskettavissa Newtonin mekaniikan avulla, jos systeemin yksityiskohdat ja sen alkutila eli ravistustapa tunnettaisiin äärettömän tarkasti. Saattaisi olla mahdollista rakentaa kone, joka näennäisesti ravistaa purkkia tarmokkaasti, mutta joka todellisuudessa pystyy säätämään, kumpi lopputulos saadaan. Sivuutamme nämä käsitteelliset vaikeudet.

Taas on luonnollista ajatella, että eri ravistusten jälkeiset lopputulokset ovat keskenään riippumattomia, koska arkijärjen mukaan tieto yhden ravistuksen lopputuloksesta ei voi vaikuttaa toisen ravistuksen lopputuloksen todennäköisyysjakaumaan.

Tällä tavalla päädymme yhteispistetodennäköisyysfunktioon

$$f(\mathbf{y};\theta) = \theta^{y_1}(1-\theta)^{1-y_1}\cdots\theta^{y_n}(1-\theta)^{1-y_n} = \theta^{t(\mathbf{y})}(1-\theta)^{n-t(\mathbf{y})}, \qquad (2.5)$$

jossa jälleen $t(\mathbf{y}) = \sum_{i=1}^{n} y_i$. Parametriavaruudeksi on luontevinta valita avoin väli $(0,1)$, sillä koejärjestely ei olisi mielekäs elleivät molemmat lopputulokset olisi mahdollisia. Tämän sijasta voimme pitää parametriavaruutena myös suljettua väliä $[0,1]$.

2.4 Binomikoe

Molemmat esimerkit ovat erikoistapauksia ns. binomikokeesta:

- Kyseessä on toistokoe, jossa tiettyä koetta toistetaan samanlaisissa olosuhteissa n kertaa; toistojen lukumäärä on tunnettu.
- Kussakin kokeessa erotetaan kaksi tulosvaihtoehtoa, joille voidaan antaa nimet onnistuminen ($Y_i = 1$) ja epäonnistuminen ($Y_i = 0$).
- Peräkkäisten toistokokeiden tulokset oletetaan toistaan riippumattomiksi, kun koetta kuvaava parametrin arvo on kiinnitetty.

Tällaisessa tilanteessa satunnaismuuttujien $Y_1,\ldots,Y_n$ yhteisjakaumalla on yptf

$$f(\mathbf{y};p) = p^{y_1}(1-p)^{1-y_1}\cdots p^{y_n}(1-p)^{1-y_n} = p^{t(\mathbf{y})}(1-p)^{n-t(\mathbf{y})},$$

jossa

$$t(\mathbf{y}) = \sum_{i=1}^{n} y_i$$

on onnistumisten lukumäärä (ykkösten lukumäärä) vektorissa $\mathbf{y}$, ja $0 \le p \le 1$ on onnistumistodennäköisyys (ykkösen todennäköisyys) yhdessä kokeessa. Pallot kulhossa -esimerkissä $p = \theta/N$, mutta nasta purkissa -esimerkissä oli $p = \theta$.

Tällaisessa tilanteessa täydellisen tulospäiväkirjan $(y_1, y_2, \ldots, y_n)$ sijasta usein raportoidaan ainoastaan onnistumisten lukumäärä

$$x = t(\mathbf{y}) = \sum_{i=1}^{n} y_i$$

kertomatta, missä järjestyksessä onnistumiset ja epäonnistumiset sattuivat. Jos onnistumisten lukumäärää pidetään satunnaismuuttujana ts. jos käsitellään satunnaismuuttujaa

$$X = t(\mathbf{Y}) = \sum_{i=1}^{n} Y_i,$$

niin tällöin X noudattaa tunnetusti *binomijakaumaa* parametreilla n ja p, jossa n on toistojen lukumäärä (tai otoskoko), ja $0 \le p \le 1$ on onnistumistodennäköisyys (ykkösen todennäköisyys) yhdessä kokeessa. Lyhyemmin merkittynä

$$X \sim \mathrm{Bin}(n,p).$$

Binomijakauman pistetodennäköisyysfunktio on

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{2.6}$$

Tästä näkökulmasta ainoa oleellinen ero näiden kahden esimerkin välillä on se, että pallot kulhossa -esimerkissä parametriavaruus on diskreetti, mutta nasta purkissa -esimerkissä parametriavaruus on jatkuva.

2.5 Kaksi lähestymistapaa

Parametrisessa mallissa havaintoja vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma tunnetaan täysin, jos mallin $f(\mathbf{y}; \theta)$ parametrin θ arvo tunnetaan, mutta tilastollisessa päättelyssä θ on tuntematon luku. Tämän takia ensimmäisenä pyrkimyksenä on arvioida eli estimoida parametrin θ arvoa havaitun aineiston $\mathbf{y}$ perusteella, ja yrittää vielä kuvailla tähän arvioon liittyvää epävarmuutta.

Historiallisesti varhaisempi lähestymistapa tähän ongelmaan tunnetaan nimellä bayesiläinen päättely. Sen perusajatuksen esitti pastori Thomas Bayes (n. 1701–1761) 1760-luvulla julkaistussa artikkelissa. Samoihin aikoihin matemaatikko Laplace (1749–1827) kehitteli ja popularisoi tätä ajattelutapaa. 1800-luvulla bayesiläinen päättely oli ainoa yleisesti tunnettu tilastollisen päättelyn periaate, joskin periaatteeseen viitattiin siihen aikaan termillä käänteinen todennäköisyys (engl. *inverse probability*).

1920-luvulla englantilainen geneetikko ja tilastotieteilijä R. A. Fisher (1890–1962) kritisoi erittäin voimakkaasti edeltäjiensä menetelmiä, ja käytännössä perusti frekventistisen päättelyn (eli ns. klassisen tai ortodoksisen tilastotieteen) esittelemällä joukon menetelmiä, joilla silloiset empiirisen tieteen tutkimusongelmat saatiin kätevästi ratkaistua. Fisherin vaikutuksen ansiosta bayesiläinen lähestymistapa unohtui lähes kokonaan.

Bayesiläinen lähestymistapa alkoi tulla uudestaan suosituksi vasta 1980-luvun loppupuolelta lähtien. Uusi nousu perustui suurelta osin uusiin laskentamenetelmiin sekä siihen, että tietokoneiden käyttö alkoi niihin aikoihin tulla jokapäiväiseksi.

2.5.1 Frekventistinen lähestymistapa

Frekventistisessä lähestymistavassa parametri θ on tuntematon, mutta kiinteä (eli ei-satunnainen) luku. Siitä tiedetään ainoastaan se, missä joukossa eli parametriavaruudessa sen arvot voivat olla.

Frekventistisessä päättelyssä *tilastollinen malli* koostuu satunnaisvektorin $\mathbf{Y}$ jakauman yptnf:stä tai ytf:stä $f(\mathbf{y}; \theta)$ sekä parametriavaruudesta Θ. Se on siis jakaumien

$$\{f(\mathbf{y}; \theta) : \theta \in \Theta\}$$

modostama perhe (termi perhe tarkoittaa samaa asiaa kuin termi joukko).

Frekventistisessä lähestymistavassa satunnaisuus viittaa aina siihen, että mikäli aineiston keruuta voitaisiin toistaa täsmälleen samoissa olosuhteissa, niin saatavat tulokset voisivat olla erilaisia. Toisin sanoen frekventistisessä päättelyssä satunnaisuus liittyy siihen, että havaitun aineiston $\mathbf{y}$ sijasta ajatellaan sitä vastaavaa satunnaisvektoria $\mathbf{Y}$ ja sen jakaumaa.

Frekventistisessä päättelyssä tutkitaan esimerkiksi seuraavia erityiskysymyksiä.

Piste-estimointi. Parametriavaruudesta pitää aineiston perusteella valita yksi arvo, jota pidetään hyvänä arvauksena parametrin todelliselle arvolle.

Väliestimointi. Parametriavaruudesta pitää rajata sellainen väli (tai joukko), jonka (tietyssä mielessä) luotetaan sisältävän oikean parametrin arvon. Tälläisen luottamusvälin avulla pyritään kuvaamaan piste-estimoinnissa saatavaa tarkkuutta.

Hypoteesintestaus. Pyritään päättämään, onko aineisto sopusoinnussa tilanteessa asetetun hypoteesin kanssa vai ei.

Mallin sopivuuden ja riittävyyden arviointi. Astutaan parametrisen mallin ulkopuolelle, ja tutkitaan, onko analyysissä käytetty malli, eli jakaumaperhe $\mathbf{y} \mapsto f(\mathbf{y};\theta), \theta \in \Theta$ lainkaan sopiva kuvaaman todellista havaittua aineistoa.

2.5.2 Bayesiläinen lähestymistapa

Bayesiläisessä lähestymistavassa myös parametri tulkitaan satunnaismuuttujaksi. Edellä käsitelty aineistoa vastaavan satunnaisvektorin jakauma $f(\mathbf{y};\theta)$ ymmärretään satunnaisvektorin $\mathbf{Y}$ ehdolliseksi jakaumaksi, kun parametrilla on arvo θ. Sille käytetään ehdollisen jakauman merkintää $f(\mathbf{y} \mid \theta)$. Kaikki koetilanteeseen liittyvä taustatieto pyritään esittämään parametrin priorijakaumana, joka on todennäköisyysjakauma parametriavaruudessa. Priorijakauman ajatuksena on esittää kvantitatiivisesti tutkijan epävarmuus parametrin oikeasta arvosta ennen (lat. *a priori*) kuin havaintoa on tehty.

Bayesiläisessä lähestymistavassa *tilastollinen malli* koostuu ehdollisesta jakaumasta $f(\mathbf{y} \mid \theta)$ sekä priorijakaumasta.

Priorijakauma ja havaintovektorin $\mathbf{Y}$ ehdollinen jakauma määräävät näiden kahden satunnaissuureen yhteisjakauman, ja bayesiläisessä päättelyssä näistä kahdesta tiedosta sitten siirrytään parametrin posteriorijakaumaan eli parametrin ehdolliseen jakaumaan, kun tiedetään, että $\mathbf{Y}$ on saanut arvon $\mathbf{y}$. Posteriorijakauma määräytyy periaatteessa automaattisesti todennäköisyyslaskennan sääntöjen avulla, mutta käytännössä sen ominaisuuksia joudutaan usein selvittämään raskaiden laskujen avulla.

Posteriorijakauma esittää kvantitatiivisesti tutkijan epävarmuuden parametrin arvosta, kun havainto otetaan huomioon. Usein myös bayesiläisessä päättelyssä lasketaan piste-estimaatteja ja väliestimaatteja, vaikka ne ovatkin vain eräitä (varsin köyhiä) tapoja kuvailla posteriorijakaumaa.

2.5.3 Yhteenveto

- Frekventistisessä päättelyssä mallin parametri on kiinteä mutta tuntematon. Lähestymistapa perustuu siihen ajatteluun, että havaitun aineiston sijasta tarkastellaan sitä vastaavaa satunnaisvektoria $\mathbf{Y}$ ja sen jakauman perusteella johdettuja jakaumia.
- Bayesiläisessä päättelyssä parametria pidetään satunnaisena, mutta aineistoa kiinteänä. Kaikki laskut ehdollistetaan käyttämällä sitä tietoa, että satunnaisvektori $\mathbf{Y}$ on saanut arvokseen havaitut arvot $\mathbf{y}$.

Luku 3

Piste-estimointi

Tarkastelemme frekventististä tilastollista mallia eli jakaumaperhettä

$$\{f(\mathbf{y};\theta), \theta \in \Theta\}$$

sekä aineistoa, jonka ajattelemme generoituneen tästä mallista eli jostakin tähän perheeseen kuuluvasta jakaumasta. Tässä luvussa esitellään menetelmiä, joilla tuntemattoman parametrin "todellista" arvoa voidaan arvioida eli estimoida. Tämä tarkoittaa sitä, että parametriavaruudesta valitaan yksi arvo $\hat{\theta}$, joka on (jonkin kriteerin mielessä) paras arvaus parametrin todelliselle arvolle. Ts. tarkasteltavasta jakaumaperheestä valitaan estimaattia $\hat{\theta}$ vastaava jakauma $\mathbf{y} \mapsto f(\mathbf{y};\hat{\theta})$, joka mielestämme paras arvaus sillä jakaumalle, joka havainnot tuotti.

Sana *todellinen* laitettiin yllä lainausmerkkeihin hyvästä syystä. Saattaa olla, että havainnot on tuottanut sellainen prosessi, jota analyysissä käyttämämme malli $f(\mathbf{y};\theta)$ ei kuvaa hyvin. Kuuluisaa tilastotieteilijää George E. P. Boxia lainaten

> *All models are wrong, but some are useful.*

Voimme olla aivan varmoja parametrisen mallin oikeellisuudesta vain harvoissa tapauksissa, kuten silloin, jos olemme aineiston simuloineet tietokoneella ko. parametrisesta mallista. Tällaisessa tapauksessa parametrin todellinen arvo on se arvo, jota käytettiin simuloinnissa.

3.1 Parametri ja tunnusluku

Sanaa parametri voi tarkoittaa tilastotieteessä eri yhteyksissä eri asioita. Tähän asti sillä on tarkoitettu sitä parametrisessa mallissa $f(\mathbf{y};\theta)$ esiintyvää lukua (tai luvuista koostuvaa vektoria) θ, jonka tunteminen kiinnittäisi havaintosatunnaisvektorin $\mathbf{Y}$ jakauman. Toisaalta sana parametri voi tarkoittaa mitä tahansa vektorin $\mathbf{Y}$ jakauman ominaisuutta kuvaavaa lukua. Pallot kulhossa -esimerkissä saattaisimme vaikkapa olla kiinnostuneita yksittäisen heiton 0/1-esityksen Y_i odotusarvosta tai varianssista, jotka ovat

$$EY_i = \frac{\theta}{N}, \qquad \operatorname{var} Y_i = \frac{\theta}{N}\left(1 - \frac{\theta}{N}\right).$$

Tämän sijasta voisimme olla kiinnostuneita summan $X = t(\mathbf{Y}) = Y_1 + \cdots + Y_n$ odotusarvosta ja varianssista

$$EX = n\,\frac{\theta}{N}, \qquad \operatorname{var} X = n\,\frac{\theta}{N}\left(1 - \frac{\theta}{N}\right).$$

Kaikkia näitä suureita voidaan kutsua parametreiksi. Parametri on yleisesti ottaen jokin mallin parametrista θ riippuva lauseke $\tau = k(\theta)$. Parametreja merkitään (pääsääntöisesti) kreikkalaisilla kirjaimilla.

Parametrista käytetään myös nimitystä populaatioparametri. Tällöin ajatellaan, että aineisto on (jollakin menetelmällä muodostettu) otos joko jostakin äärellisestä populaatiosta tai jostakin (kuvitteellisesta) äärettömästä populaatiosta. Estimoinnin tavoitteena on tehdä johtopäätöksiä ko. populaatiosta (ts. populaatioparametreista) havaintojen avulla. Tällöin soveltajan tulee tarkoin miettiä, mitä populaatiota havaintoaineisto edustaa, eli mihin populaatioon tilastolliset johtopäätökset voidaan yleistää.

Tunnusluku (engl. *statistic*) tarkoittaa mitä tahansa lukua, joka voidaan laskea aineistosta. Binomikokeessa onnistumisten lukumäärä $t(\mathbf{y}) = \sum_{i=1}^{n} y_i$ on eräs tunnusluku. Kaikki tunnusluvut voidaan esittää kaavalla $t(\mathbf{y})$ jossa funktio t valitaan kulloisenkin tilanteen mukaan, ja funktio t ei saa riippua mistään mallin tuntemattomasta parametrista.

3.2 Estimaatti, estimaattori ja otantajakauma

Määritelmä 3.1 (Estimaatti)**.** Joitakin tunnuslukuja käytetään parametrien arvioina, jolloin niitä kutsutaan vastaavien parametrien *estimaateiksi*.

Nasta purkissa -esimerkissä onnistumistodennäköisyyttä θ tavallisesti arvioitaisiin laskemalla onnistumisten suhteellinen osuus n kokeessa, eli

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.1}$$

Estimaatteja on tapana merkitä kuten edellä tehtiin, eli laittamalla hattu vastaavan parametrin päälle. Jos tarjolla on monta erilaista estimaattia samalle parametrille, niin ne voidaan erottaa toisistaan esimerkiksi lisäämällä merkintöihin ala- tai yläindeksejä.

Eräs minimaalinen järkevyysvaatimus estimaatille on se, että mallin parametrin θ estimaatin $\hat{\theta}$ pitää kuulua parametriavaruuteen Θ. Vastaavasti parametrin $\tau = k(\theta)$ estimaatin $\hat{\tau}$ pitää kuulua joukkoon

$$\{k(\theta) : \theta \in \Theta\}.$$

Nasta purkissa -esimerkin estimaatille (3.1) tämä toteutuu automaattisesti, mikäli parametriavaruudeksi on valittu $[0, 1]$. Mikäli parametriavaruudeksi valitaan avoin väli $(0, 1)$, niin estimaatti (3.1) ei täytä tätä minimaalista vaatimusta, mikäli nasta ei päädy kertaakaan selälleen (jolloin $\sum_i y_i = 0$) tai mikäli nasta ei päädy kertaakaan kyljelleen (jolloin $\sum y_i = n$).

Pallot kulhossa -esimerkissä onnistumisten suhteellista osuutta (3.1) voitaisiin ehdottaa onnistumistodennäköisyyden $\phi = \theta/N$ estimoimiseen mutta tällöin törmättäisiin siihen ongelmaan, että tämän parametrin ϕ arvot kuuluvat

mallissa joukkoon

$$\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\},$$

mutta sen estimaatti $\hat{\phi}$ voi saada arvoja joukosta

$$\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\},$$

eikä näillä joukoilla välttämättä ole edes kovin montaa yhteistä alkiota. Tämä ongelma pitäisi käytännössä kiertää pyöristämällä suhteellinen osuus jollakin tavalla diskreettiin parametriavaruuteen.

Frekventistisessä päättelyssä tunnusluvun $t(\mathbf{y})$ lisäksi tarkastellaan sitä vastaavaa satunnaismuuttujaa $t(\mathbf{Y})$. Tällöin tunnuslukua ei lasketa havaitusta aineistosta, vaan se lasketaan aineistoa vastaavasta satunnaisvektorista $\mathbf{Y}$, jolla oletetaan olevan jokin todennäköisyysjakauma. Niin kauan kuin pysytään mallin $\{f(\mathbf{y};\theta) : \theta \in \Theta\}$ puitteissa (ja joskus on mielekästä laajentaa tarkastelu mallin ulkopuolelle), oletetaan että satunnaisvektorilla $\mathbf{Y}$ on todellista parametrinarvoa θ vastaava todennäköisyysjakauma.

Määritelmä 3.2 (Otantajakauma)**.** Satunnaismuuttujan $t(\mathbf{Y})$ jakaumaa kutsutaan tämän tunnusluvun otantajakaumaksi (engl. *sampling distribution*). Oletamme, että $\mathbf{Y}$ noudattaa jakaumaa $f(\mathbf{y};\theta)$ todellisella parametrinarvolla θ.

Termissä otantajakauma on taustalla ajatus otannan tai aineiston keruun toistamisesta. Jos aineiston keruu voitaisiin toistaa samoissa olosuhteissa riippumattomasti r kertaa, ja saataisiin aineistot $\mathbf{y}_1, \ldots, \mathbf{y}_r$ (jossa kukin $\mathbf{y}_i$ on n-vektori), niin tällöin arvot $t(\mathbf{y}_1), \ldots t(\mathbf{y}_r)$ olisivat otos satunnaismuuttujaksi ymmärretyn tunnusluvun $t(\mathbf{Y})$ jakaumasta. Tämä ajatus voidaan toteuttaa konkreettisesti tietokoneella. Annetaan parametrisessa mallissa parametrille θ jokin lukuarvo, ja simuloidaan otos $\mathbf{y}_1, \ldots, \mathbf{y}_r$ jakaumasta $f(\mathbf{y};\theta)$. Tällaisia simulointimenetelmiä on saatavilla lukuisille yhteisjakaumille $f(\mathbf{y};\theta)$.

Teen kaksi frekventististä päättelyä koskevaa huomautusta.

- Parametri θ on frekventistisessä päättelyssä kiinteä mutta tuntematon luku, jolla ei ole todennäköisyysjakaumaa.
- Frekventistisessä päättelyssä tarkastellaan parametriavaruudessa määriteltyjä jakauma, mutta ne ovat aina jonkin tunnusluvun otantajakaumia.

Frekventistisessä tilastotieteessä erityisen kiinnostava asia on estimaattorin otantajakauma. Sana estimaattori tarkoittaa sitä, että estimaatin ei ajatella olevan konkreettinen luku, vaan sen ajatellaan olevan satunnaismuuttuja. Estimaattia $\hat{\theta} = t(\mathbf{y})$ vastaa estimaattori $t(\mathbf{Y})$, joka on satunnaismuuttuja. Voimme merkitä sitä myös kaavalla

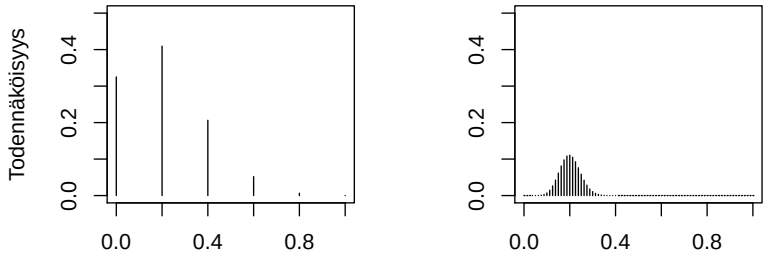
$$\hat{\theta}(\mathbf{Y}) = t(\mathbf{Y}).$$

Mallimme puitteissa estimaatti $t(\mathbf{y})$ on estimaattorin $\hat{\theta}(\mathbf{Y})$ havaittu arvo, sillä $\mathbf{y} = \mathbf{Y}(\omega^{\text{act}})$ (ks. kaava (2.1)).

Nasta purkissa -esimerkissä estimaattia

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Kuva 3.1 Estimaattorin "onnistumisten suhteellinen osuus binomikokeessa" otantajakauma, kun $\theta = 0.2012$ ja $n = 5$ (vasemmalla) ja $n = 80$ (oikealla).



vastaa estimaattori

$$\hat{\theta}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{3.2}$$

jonka otantajakauma on skaalausta vaille sama kuin tunnusluvun $\sum Y_i$ jakauma, joka puolestaan on binomijakauma $\mathrm{Bin}(n, \theta)$. Estimaattorin (3.2) otantajakauma on tällä perusteella

$$P_\theta(\hat{\theta}(\mathbf{Y}) = \frac{k}{n}) = \binom{n}{k} \theta^k \, (1-\theta)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Kuvassa 3.1 esitetään tämän diskreetin otantajakauman pistetodennäköisyysfunktio kahdelle erilaiselle otoskoolle n.

3.3 Todennäköisyyslaskennan tietoja

Palautetaan tässä välissä mieleen joitakin tietoja todennäköisyyslaskennasta.

Jos X on satunnaismuuttuja, ja a on vakio, niin skaalatun satunnaismuuttujan $a\,X$ odotusarvo ja varianssi ovat

$$E(a\,X) = a\,EX, \qquad \mathrm{var}(a\,X) = a^2 \,\mathrm{var}\, X. \tag{3.3}$$

Jos X_1 ja X_2 ovat satunnaismuuttujia, niin niiden summan odotusarvo on odotusarvojen summa,

$$E(X_1 + X_2) = EX_1 + EX_2. \tag{3.4}$$

Jos X_1 ja X_2 ovat *riippumattomia* satunnaismuuttujia, niin niiden summan tai erotuksen varianssi saadaan laskemalla yhteen muuttujien varianssit, eli

$$\mathrm{var}(X_1 \pm X_2) = \mathrm{var}\, X_1 + \mathrm{var}\, X_2. \tag{3.5}$$

Jos $\mu = EX$, ja a on vakio, niin helpolla laskulla nähdään, että

$$E(X-a)^2 = E(X-\mu)^2 + (\mu - a)^2 = \mathrm{var}\, X + (\mu - a)^2 \tag{3.6}$$

Tšebyševin epäyhtälön mukaan mille tahansa vakiolle a

$$P(|X - a| > \epsilon) \leq \frac{E(X-a)^2}{\epsilon^2}, \qquad \text{kaikille } \epsilon > 0. \tag{3.7}$$

3.4 Otantajakauman ominaisuuksia

Binomikokeessa estimaattorin (3.2) eli onnistumisten suhteellisen osuuden (otantajakauman) odotusarvo ja varianssi ovat helppo selvittää, sillä

$$E_\theta[\hat{\theta}(\mathbf{Y})] = \frac{1}{n}\sum_{i=1}^{n} E_\theta(Y_i) = \frac{1}{n}\, n\,\theta = \theta,$$
$$\mathrm{var}_\theta[\hat{\theta}(\mathbf{Y})] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}_\theta(Y_i) = \frac{1}{n^2}\, n\,\theta\,(1-\theta) = \frac{1}{n}\,\theta\,(1-\theta)$$

Alaindeksillä θ korostetaan sitä, että satunnaisvektorilla $\mathbf{Y} = (Y_1, \ldots, Y_n)$ oletetaan olevan mallin $f(\mathbf{y}; \theta)$ mukainen jakauma. Otantajakauman varianssin määrittäminen perustui siihen mallin oletukseen, että satunnaismuuttujat Y_i ovat riippumattomia. Vaihtoehtoisesti voimme johtaa odotusarvon ja varianssin käyttämällä hyväksi tunnettuja kaavoja binomijakauman $\mathrm{Bin}(n, \theta)$ odotusarvolle ja varianssille.

Määritelmä 3.3 (Harhattomuus)**.** Jos estimaattorin odotusarvo on sama kuin parametrin todellinen arvo, eli

$$E_\theta[\hat{\theta}(\mathbf{Y})] = \theta, \quad \text{kaikilla } \theta,$$

niin sanotaan, että estimaattori $\hat{\theta}(\mathbf{Y})$ on *harhaton* (engl. *unbiased*). Muussa tapauksessa sanotaan, että estimaattori on harhainen (engl. *biased*).

Tarkemmin sanoen edellinen asia voidaan ilmaista niin, että estimaattori on *odotusarvon mielessä* harhaton; odotusarvon sijasta voisimme toki tarkastella muitakin otantajakauman keskikohtaa kuvailevia suureita, kuten mediaania tai moodia.

Määritelmä 3.4 (Harha)**.** Estimaattorin $\hat{\theta}(\mathbf{Y})$ harha on

$$\mathrm{bias}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta(\hat{\theta}(\mathbf{Y})) - \theta. \tag{3.8}$$

Mallin parametrin θ sijasta voitaisiin tarkastella myös jotakin muuta parametria $\tau = k(\theta)$ estimoivan estimaattorin $\hat{\tau}(\mathbf{Y})$ harhaa. Tämä tietenkin määritellään edellistä vastaavalla kaavalla

$$\mathrm{bias}_\theta(\hat{\tau}(\mathbf{Y})) = E_\theta(\hat{\tau}(\mathbf{Y})) - k(\theta). \tag{3.9}$$

Harha voidaan määritellä samalla kaavalla myös silloin, jos parametri on vektori.

Harhaa pidetään usein estimaattorin systemaattisena virheenä. Harhattoman estimaattorin harha on nolla koko parametriavaruudessa. Harhainen estimaattori ei kuitenkaan välttämättä ole huono estimaattori eikä harhaton estimaattori ole välttämättä hyvä estimaattori. Nasta purkissa -esimerkissä estimaattori (3.2) on harhaton.

Merkinnät alkavat tässä vaiheessa arvatenkin näyttää raskailta, joten avaan seuraavaksi niiden merkitystä estimaattorin harhan määritelmän eli kaavan (3.8)

$$\mathrm{bias}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta(\hat{\theta}(\mathbf{Y})) - \theta$$

kohdalla.

- Siinä puhutaan estimaattorista $\hat{\theta}(\mathbf{Y})$, jota siis käsitellään satunnaismuuttujana.
- Estimaattori $\hat{\theta}(\mathbf{Y})$ on funktio satunnaisvektorista $\mathbf{Y}$, joten estimaattorin jakauma riippuu satunnaivektorin $\mathbf{Y}$ jakaumasta.
- Alaindeksi θ kertoo, että satunnaisvektorin $\mathbf{Y}$ jakaumalla on yptnf tai ytf $f(\mathbf{y};\theta)$.

Näissä luentomuistiinpanoissa käytetään tällaisia pedanttisia merkintöjä, jotta lukija pystyisi kaavoista heti näkemään, mitä suureita pidetään kiinteinä ja mitä satunnaisina ja mitä jakaumia satunnaisille suureille oletetaan. Sen jälkeen, kun nämä asiat alkavat olla itsestään selviä, opiskelija voi rauhassa tiputtaa kaavoista ylimääräiset koristeet, ja kirjoittaa vaikkapa

$$\text{bias}(\hat{\theta}) = E\hat{\theta} - \theta,$$

millä tyylillä nämä asiat monessa oppikirjassa esitetään. Kirjallisuudessa ei välttämättä tehdä eroa termien estimaatti ja estimaattori välillä, vaan termi estimaatti saattaa tarkoittaa niistä kumpaa tahansa. Lisäksi merkintä $\hat{\theta}$ saattaa kontekstista riippuen tarkoittaa yhtä hyvin estimaattia tai estimaattoria.

Määritelmä 3.5 (Keskineliövirhe)**.** Estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirhe (engl. *mean squared error*) on

$$\text{mse}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta\left[(\hat{\theta}(\mathbf{Y}) - \theta)^2\right] \tag{3.10}$$

Keskineliövirhe kuvaa estimaattorin tarkkuutta: mitä pienempi keskineliövirhe, sitä tarkempia arvioita keskimäärin saadaan. Keskineliövirhe riippuu tyypillisesti voimakkaasti otoskoosta n siten, että suuremmalla otoskoolla saavutetaan pienempi keskineliövirhe.

Mikäli estimaattori on harhaton, niin sen keskineliövirhe on sama kuin sen varianssi. Helpolla laskulla (vrt. kaava (3.6)) nähdään, että keskineliövirhe voidaan esittää laskemalla yhteen estimaattorin varianssi ja sen harhan neliö, eli

$$\text{mse}_\theta(\hat{\theta}(\mathbf{Y})) = \text{var}_\theta(\hat{\theta}(\mathbf{Y})) + \Big(\text{bias}_\theta(\hat{\theta}(\mathbf{Y}))\Big)^2. \tag{3.11}$$

Keskineliövirheen sijasta usein tarkastellaan sen neliöjuurta, koska se on samalla skaalalla kuin itse estimaattori.

Määritelmä 3.6 (Keskineliövirheen neliöjuuri, RMSE)**.** Estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirheen neliöjuuri (engl. *root mean squared error*) on

$$\text{rmse}_\theta(\hat{\theta}(\mathbf{Y})) = \sqrt{\text{mse}_\theta(\hat{\theta}(\mathbf{Y}))}. \tag{3.12}$$

Estimaattien yhteydessä usein kerrotaan niiden *keskivirhe*. Tämä on yksi tapa arvioida estimointiin liittyvää epävarmuutta.

Määritelmä 3.7 (Keskivirhe)**.** Estimaatin $\hat{\theta}$ keskivirhe (engl. *standard error, s.e., se*) tarkoittaa otoksesta (jollakin järkevällä tavalla) muodostettua estimaattia vastaavan estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirheen neliöjuurelle (eli RMSE:lle).

Estimaattorin keskineliövirheen neliöjuuri (eli RMSE) riippuu yleensä jollakin tavalla parametrinarvosta θ, ja kun tähän kaavaan sijoitetaan tuntemattoman parametrin tai tuntemattomien parametrien tilalle niiden estimaatit, niin saadaan estimaatin keskivirhe. Tyypillisesti keskivirheestä puhutaan silloin, kun vastaava estimaattori on harhaton. Tällöin sen keskineliövirheen neliöjuuri on sama asia kuin estimaattorin (otantajakauman) varianssin neliöjuuri. Varianssin neliöjuuresta käytetään nimitystä keskihajonta (engl. *standard deviation*). *Harhatonta estimaattoria vastaavan estimaatin keskivirhe on kyseisen estimaattorin otantajakauman estimoitu keskihajonta.*

Mikäli keskineliövirhe (tai sen neliöjuuri) suppenee kohti nollaa, kun otoskoko n kasvaa rajatta, niin tällöin nähdään Tšebyševin epäyhtälön (3.7) avulla, että estimaattori $\hat{\theta}(\mathbf{Y})$ suppenee stokastisesti (engl. *converges in probability*) kohti parametrin todellista arvoa, eli

$$\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta.$$

Stokastinen suppeneminen tarkoittaa sitä, että kaikilla $\epsilon > 0$ pätee, että

$$P_\theta\{|\hat{\theta}(\mathbf{Y}) - \theta| \geq \epsilon\} \to 0, \qquad \text{kun } n \to \infty. \tag{3.13}$$

Määritelmä 3.8 (Tarkentuvuus)**.** Jos $\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta$ kaikilla $\theta \in \Theta$, niin sanotaan, että estimaattori $\hat{\theta}(\mathbf{Y})$ on *tarkentuva* (engl. *consistent*).

(Tarkemmin sanoen näin määritellään estimaattorin heikko tarkentuvuus.) Tarkentuvuus tarkoittaa sitä, että otoskoon kasvaessa estimaattorin otantajakauma keskittyy yhä tiiviimmin ja tiiviimmin parametrin todellisen arvon ympärille.

Nasta purkissa -esimerkissä estimaattorin (3.2) keskineliövirhe on harhattomuuden ansiosta sama kuin sen varianssi, joten

$$\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y})) = \mathrm{var}_\theta(\hat{\theta}(\mathbf{Y})) = \frac{1}{n}\,\theta\,(1-\theta),$$

ja koska tämä suppenee otoskoon kasvaessa kohti nollaa, on estimaattori tarkentuva.

Monimutkaisissa tapauksissa todennäköisyyslaskennan taitomme eivät aina riitä estimaattorin otantajakauman ominaisuuksien selvittämiseen. Tällöin niitä voidaan yrittää selvittää tietokonesimuloinnin avulla.

Frekventistisessä tilastotieteessä erilaisia estimaattoreja verrataan keskenään niiden otantajakaumien ominaisuuksien (kuten esimerkiksi harhan ja varianssin) avulla. Kun estimaatti sitten lasketaan aineistosta, niin (epämuodollisesti) ajatellaan, että kyseinen estimaatti on tarkka, mikäli vastaavalla estimaattorilla on suotuisa otantajakauma (esim. pieni harha ja pieni varianssi).

3.5 Uskottavuusfunktio

Kun aineisto $\mathbf{y}$ on havaittu, ja havaittua arvoa käytetään funktion $f(\mathbf{y};\theta)$ ensimmäisenä argumenttina, niin parametriavaruudella määriteltyä funktiota

$$\theta \mapsto f(\mathbf{y};\theta)$$

kutsutaan *uskottavuusfunktioksi* (engl. *likelihood function*). Sitä merkitään

$$L(\theta) = f(\mathbf{y};\theta).$$

Joskus tahdotaan kirjata näkyviin, että uskottavuusfunktio riippuu myös aineistosta $\mathbf{y}$, ja tällöin voidaan käyttää merkintää

$$L(\theta;\mathbf{y}) = f(\mathbf{y};\theta).$$

Haluttaessa voidaan sanoa tarkemmin, että kyseessä on havaintoa $\mathbf{y}$ vastaava parametrin θ uskottavuusfunktio.

Huomaa, että uskottavuusfunktion yhteydessä θ on vapaa muuttuja, eikä tarkoita parametrin todellista arvoa. Kuten aikaisemmin todettiin, tälläinen symbolien väärinkäyttö tarkoittamaan erilaisissa yhteyksissä aivan erilaisia asioita on tilastotieteen merkinnöille tyypillistä, eikä se huolellisesti käytettynä ja tulkittuna aiheuta sekaannusta.

Funktio

$$\mathbf{y} \to f(\mathbf{y};\theta)$$

eli lauseke $f(\mathbf{y};\theta)$ ymmärrettynä argumentin $\mathbf{y}$ funktiona kiinteällä θ on satunnaisvektorin $\mathbf{Y}$ yhteistiheysfunktio tai yhteispistetodennäköisyysfunktio. Tästä poiketen uskottavuusfunktiossa argumentti $\mathbf{y}$ kiinnitetään sijoittamalla siihen havaitut arvot. Näin saatua lauseketta tarkastellaan parametrin funktiona eli uskottavuusfunktio on parametriavaruudella määritelty kuvaus, joka saa pisteessä θ arvon

$$L(\theta) = f(\mathbf{y};\theta), \qquad \theta \in \Theta.$$

jossa yptnf:n tai ytf:n argumentin $\mathbf{y}$ arvoksi on kiinnitetty havaitut arvot. Uskottavuusfunktio ei ole pistetodennäköisyysfunktio eikä tiheysfunktio.

Esimerkki 3.1. Oletetaan pallot kulhossa -esimerkissä (jakso 2.2), että kulhossa on $N = 5$ palloa ja että nostot tehdään palauttaen ja että tulokset ovat $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$. Tällöin valkoisten pallojen lukumäärä $n = 7$ nostossa (eli onnistumisten lukumäärä) on 2, ja uskottavuusfunktio on kaavan (2.4) mukaan

$$L(\theta) = \left(\frac{\theta}{5}\right)^2 \left(1 - \frac{\theta}{5}\right)^5, \qquad \theta = 0, 1, 2, 3, 4 \text{ tai } 5,$$

jossa onnistumistodennäköisyys on θ/N, joka on valkoisten pallojen suhteellinen osuus kulhossa. △

Usein kannattaa uskottavuusfunktion sijasta tarkastella sen logaritmia.

Määritelmä 3.9 (Logaritminen uskottavuusfunktio)**.** Uskottavuusfunktion logaritmia

$$\ell(\theta) = \log L(\theta)$$

kutsutaan logaritmiseksi uskottavuusfunktioksi tai uskottavuusfunktion logaritmiksi tai log-uskottavuusfunktioksi (engl. *log-likelihood*). Tässä log tarkoittaa luonnollista logaritmia.

Silloin, kun siirrytään uskottavuusfunktiosta $L(\theta)$ logaritmiseen uskottavuusfunktioon $\ell(\theta) = \log L(\theta)$ tehdään tavallisesti se oletus, että $L(\theta) > 0$ koko parametriavaruudessa, jolloin $\ell(\theta)$ on hyvin määritelty reaalifunktio: $\log(0)$ ei ole reaaliluku. Vaihtoehtoinen tapa selvitä tästä pulmasta on sopia, että $\log(0) = -\infty$, joka on pienempi kuin mikään reaaliluku.

Logaritmointi on kätevää monesta syystä. Jos uskottavuusfunktio on tulomuotoa (2.2), niin logaritmin otto muuttaa sen summaksi, sillä

$$\log(\prod_{i=1}^{n} f_{Y_i}(y_i;\theta)) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i;\theta)).$$

Tässä sovellettiin tuttua kaavaa

$$\log(ab) = \log(a) + \log(b), \qquad \text{kun } a > 0 \text{ ja } b > 0.$$

Tietokoneella laskettaessa logaritmointi on tärkeää, sillä uskottavuusfunktiossa esiintyvät tulon termit ovat usein erittäin pieniä lukuja, jolloin itse uskottavuusfunktion arvoksi saattaa tietokoneohjelmassa tulla tasan nolla, vaikka kyseessä olisi aidosti positiivinen luku. Logaritmin ottaminen uskottavuusfunktiosta riittää yleensä ratkaisemaan tämän ongelman.

Binomikokeessa uskottavuusfunktion tai sen logaritmin arvo voidaan laskea missä tahansa parametriavaruuden pisteessä heti kun tiedetään onnistumisten lukumäärä n kokeessa ilman, että tarvitsee tietää täydellistä tulospäiväkirjaa $(y_1, \ldots, y_n)$. Tämä ilmaistaan sanomalla, että onnistumisten lukumäärä on tyhjentävä tunnusluku. Myös onnistumisten suhteellinen osuus k/n tai moni muu vastaava suure on binomikokeessa tyhjentävä tunnusluku.

Määritelmä 3.10 (Tyhjentävä tunnusluku)**.** Tunnusluku $t(\mathbf{y})$ on tyhjentävä, mikäli uskottavuusfunktion arvo voidaan laskea yksinomaan sen perusteella.

Edellisessä määritelmässä sallitaan myös se tapaus, jossa t on vektoriarvoinen funktio. Esimerkiksi koko aineisto $\mathbf{y}$ on missä tahansa mallissa (triviaalisti) tyhjentävä tunnusluku.

3.6 Suurimman uskottavuuden estimaatti

Frekventistisessä tilastotieteessä parametria θ pidetään tuntemattomana vakiona, josta tiedetään vain, missä joukossa (eli parametriavaruudessa) sen arvot voivat olla. Parametria voidaan estimoida eli arvioida erilaisilla menetelmillä.

Tunnetuin estimointiperiaate on ns. *suurimman uskottavuuden*, eli SU-periaate (engl. *maximum likelihood, ML*), jonka mukaan parametrin parhaana estimaattina pidetään sitä parametriavaruuden arvoa $\hat{\theta}$, joka maksimoi uskottavuusfunktion. Sitä kutsutaan suurimman uskottavuuden estimaatiksi (eli SU-estimaatiksi) (engl. *maximum likelihood estimate, ML estimate, MLE*). Tämä ajatus voidaan esittää kaavalla

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta). \tag{3.14}$$

Merkintä $\arg\max L(\theta)$ tarkoittaa lausekkeen $L(\theta)$ maksimoivaa argumenttia (ts. maksimipistettä). Sen sijaan merkintä $\max L(\theta)$ tarkoittaisi lausekkeen $L(\theta)$

maksimiarvoa. Kun näitä merkintöjä käytetään, niin tällöin hiljaisesti oletetaan, että parametriavaruudessa on olemassa yksikäsitteinen maksimipiste $\hat{\theta}$, jolle

$$L(\hat{\theta}) \geq L(\theta), \quad \text{kaikille } \theta \in \Theta.$$

Mikäli aineistoa vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma on diskreetti, niin SU-estimaatti on se parametrialueen piste, joka tekee havaitun aineiston (mallin puitteissa) mahdollisimman todennäköiseksi, eli

$$P_{\hat{\theta}}(\mathbf{Y} = \mathbf{y}) \geq P_{\theta}(\mathbf{Y} = \mathbf{y}), \qquad \text{kaikilla } \theta \in \Theta.$$

Jatkuvan yhteisjakauman tapauksessa tulkinta on samantapainen: SU-estimaatti on se parametriarvo, joka maksimoi yhteistiheysfunktion arvon laskettuna aineistolle $\mathbf{y}$.

Esimerkki 3.2. (Jatkoa esimerkille 3.1, pallot kulhossa) Valkoisten pallojen lukumäärä θ on yksi luvuista $0, 1, 2, 3, 4$ tai 5, ja uskottavuusfunktio on

$$\begin{aligned} L(\theta) &= \left(\frac{\theta}{5}\right)^2 \left(1 - \frac{\theta}{5}\right)^5 \\ &= \begin{cases} 0, & \text{jos } \theta = 0, \\ 1024/5^7, & \text{jos } \theta = 1, \\ 972/5^7, & \text{jos } \theta = 2, \\ 288/5^7, & \text{jos } \theta = 3, \\ 16/5^7, & \text{jos } \theta = 4, \\ 0, & \text{jos } \theta = 5. \end{cases} \end{aligned}$$

Havaintojen valossa voimme sulkea pois arvot $\theta = 0$ ja $\theta = 5$, koska kulhosta ei voitaisi nostaa valkoisia (mustia) palloja, jos niitä ei siellä alunperin lainkaan olisi. Voimme myös sanoa, että arvo $\theta = 3$ on uskottavampi kuin arvo $\theta = 4$, koska $L(3) > L(4)$. Todennäköisyys poimia valkoinen pallo kaksi kertaa seitsemässä nostossa on suurempi, mikäli $\theta = 3$ kuin siinä tapauksessa, että $\theta = 4$. Kaikista uskottavin arvo eli SU-estimaatti on $\hat{\theta} = 1$. △

Varoitus. SU-estimaatti $\hat{\theta}$ on edellisessä esimerkissä se arvo, joka tekee havainnot (mallin puitteissa) mahdollisimman todennäköisiksi. Sen sijaan olisi vakava väärinkäsitys väittää, että $\hat{\theta}$ eli uskottavin parametrin olisi parametrin todennäköisin arvo. Frekventistisen tilastotieteen puitteissa tällainen lausuma on mieltä vailla, koska parametrin arvoa koskevia todennäköisyyksiä ei frekventistisessä mallissa ole määriteltynä. Juuri tästä syystä Fisher otti käyttöön termin *uskottavuus*.

Jos mallista löytyy tyhjentävä tunnusluku, niin tällöin SU-estimaatti riippuu aineistosta vain tyhjentävän tunnusluvun arvon kautta.

Esimerkki 3.3. (Jatkoa esimerkille 3.1) Olkoon palloja yhteensä $N = 5$ ja tehdään nostoja palauttaen $n = 7$ kertaa. Jos k on onnistumisten lukumäärä eli

$k = y_1 + y_2 + \cdots + y_7$, niin tällöin SU-estimaatti saadaan kaavalla

$$\hat{\theta} = \begin{cases} 0, & \text{jos } k = 0, \\ 1, & \text{jos } k = 1 \text{ tai } k = 2, \\ 2, & \text{jos } k = 3, \\ 3, & \text{jos } k = 4, \\ 4, & \text{jos } k = 5 \text{ tai } k = 6, \\ 5, & \text{jos } k = 7. \end{cases}$$

Tämä saatiin selville laskemalla uskottavuusfunktio (tietokoneella) kaikissa näissä tapauksissa, ja katsomalla missä maksimipiste kulloinkin on. △

Koska logaritmi on aidosti kasvava funktio, on uskottavuusfunktiolla $L(\theta)$ ja logaritmisella uskottavuusfunktiolla $\ell(\theta)$ samat maksimipisteet. Tämän takia SU-estimaatti voidaan yhtä hyvin määrittää kaavalla

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta). \tag{3.15}$$

Jos parametriavaruus on jatkuva, niin tavallisesti SU-estimaatti etsitään ratkaisemalla logaritmisen uskottavuusfunktion derivaatan nollakohdat. Lisäksi pitää kiinnittä huomiota (log-)uskottavuusfunktion käyttäytmiseen, kun lähestytään parametriavaruuden reunapisteitä.

3.7 SU-estimaatti binomikokeessa

Johdamme seuraavaksi SU-estimaatin kaavan binomikokeen tapauksessa silloin, kun n toistossa onnistutaan k kertaa, ja onnistumistodennäköisyys yhdessä toistossa on θ. Oletamme, että parametriavaruus Θ on joko avoin väli $(0, 1)$ tai suljettu väli $[0, 1]$.

Käsittelemme ensin sen tapauksen, jossa onnistumisten lukumäärä k on välillä $1 \leq k \leq n - 1$. Logaritminen uskottavuusfunktio on

$$\ell(\theta) = \log\left(\theta^k (1-\theta)^{n-k}\right) = k \log\theta + (n-k)\log(1-\theta),$$

joka on hyvin määritelty, kun $0 < \theta < 1$.

Ratkaisemme seuraavaksi logaritmisen uskottavuusfunktion derivaatan nollakohdat. Kun $0 < \theta < 1$, niin

$$\ell'(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta} = \frac{k - n\theta}{\theta(1-\theta)}$$

Derivaatan ainoa nollakohta on $\hat{\theta} = k/n$, ja kyseessä on maksimipiste, sillä derivaatan merkki vaihtuu siinä positiivisesta negatiiviseksi. (Nimittäjä $\theta(1-\theta)$ on positiivinen.)

Tapauksessa $k = n$ uskottavuusfunktio on

$$L(\theta) = \theta^n,$$

ja tämä on selvästi aidosti kasvava funktio välillä $(0, 1)$. Jos parametriavaruus on $[0, 1]$, niin SU-estimaatti on $\hat{\theta} = 1 = k/n$. Huomaa, että SU-estimaatti ei tässä

tapauksessa löydy derivaatan nollakohdasta, vaan parametriavaruuden reunalta. Jos parametriavaruus kuitenkin on avoin väli $(0, 1)$, niin tällöin joudumme toteamaan, että SU-estimaattia ei ole olemassa, koska uskottavuusfunktio ei saavuta missään parametriavaruuden pisteessä maksimiarvoaan.

Tapauksessa $k = 0$ nädään vastaavasti, että SU-estimaatti on $\hat{\theta} = 0 = k/n$, mikäli parametriavaruus on $[0, 1]$. Jos parametriavaruus kuitenkin on $(0, 1)$, niin SU-estimaattia ei ole olemassa.

Mikäli binomikokeessa tahdotaan käyttää SU-estimointia, niin tästä syystä on kätevää valita parametriavaruudeksi suljettu väli $[0, 1]$. Tällöin SU-estimaatti saadaan kaikissa tapauksissa kaavalla

$$\hat{\theta} = \frac{k}{n} \tag{3.16}$$

eli SU-estimaatti on onnistumisten (k) suhteellinen osuus (n toistossa).

Olemme jo edellä jaksossa 3.4 nähneet, että vastaava estimaattori on harhaton ja että sen (otantajakauman) varianssi saadaan kaavalla

$$\frac{1}{n}\,\theta\,(1-\theta).$$

Tämän ansiosta SU-estimaatin $\hat{\theta}$ keskivirhe voidaan laskea kaavalla

$$\sqrt{\frac{1}{n}\,\hat{\theta}\,(1-\hat{\theta})} \tag{3.17}$$

Kuvassa 3.2 esitetään binomikokeen uskottavuusfunktio ja logaritminen uskottavuusfunktio kahdella erilaisella otoskoolla. Näissä kuvissa tilanne on valittu siten, että $\hat{\theta} = k/n = 0.2$ on molemmilla otoskoilla. Huomaa, että pienellä otoskoolla uskottavuusfunktio on selvästi laakeampi kuin suurella otoskoolla. Suurella otoskoolla uskottavat parametrinarvot ovat melko kapealla välillä SU-estimaatin ympärillä, joten intuitio sanoo, että suurella otoskoolla parametrin arvosta voi tehdä tarkempia päätelmiä kuin pienellä. Tämän asian näkee myös laskemalla estimaattien keskivirheet kaavalla (3.17), jolloin otoskoolla $n = 5$ saadaan keskivirhe

$$\sqrt{\frac{1}{5} \times 0.2 \times 0.8} = 0.18$$

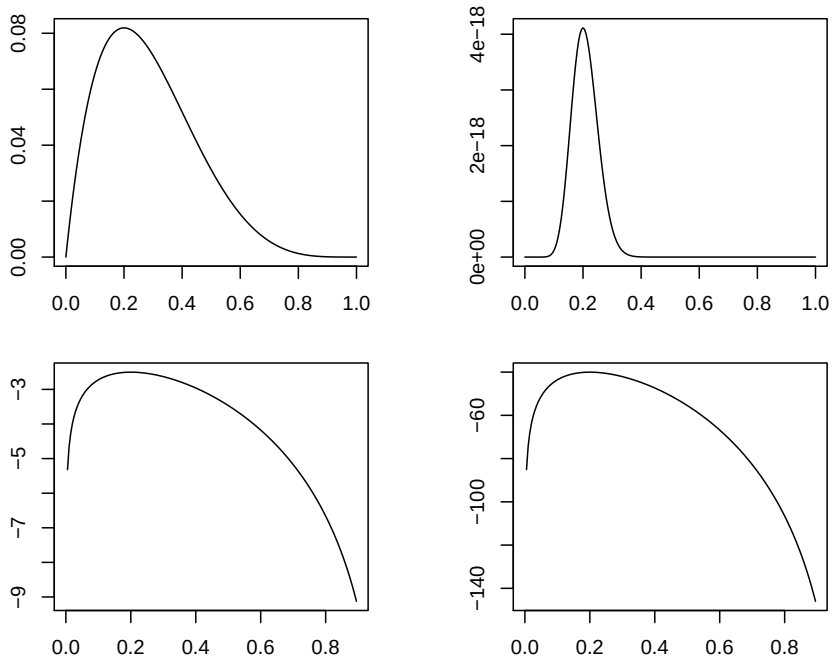
ja otoskoolla $n = 80$ keskivirhe

$$\sqrt{\frac{1}{80} \times 0.2 \times 0.8} = 0.045.$$

3.8 Normaalijakauman parametrien estimointi

Tarkastelemme tilannetta, jossa mallinnamme aineiston $\mathbf{y} = (y_1, \ldots, y_n)$ siten, että vastaavat satunnasimuuttujat $Y_1, \ldots, Y_n$ ovat satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$. Ts. oletamme, että satunnaismuuttujat Y_i ovat riippumattomia, ja kukin niistä noudattaa normaalijakaumaa $N(\mu, \sigma^2)$. Tässä $\mu \in \mathbb{R}$ ja $\sigma^2 > 0$ voivat molemmat olla tuntemattomia parametreja, tai sitten toinen niistä voi olla tunnettu vakio ja toinen tuntematon parametri.

Kuva 3.2 Uskottavuusfunktio ja logaritminen uskottavuusfunktio binomikokeessa kahdella eri otoskoolla, kun parametriavaruus on jatkuva. Vasemmalla $n = 5$ ja oikealla $n = 80$; ylempänä on uskottavuusfunktio ja alempana sen logaritmi. Molemmissa tapauksissa onnistumisten suhteellinen osuus $k/n = 0.2$. Suuremmalla otoskoolla uskottavuusfunktio ja sen logaritmi ovat selvästi terävämpihuippuisia funktioita kuin pienellä; logaritmisten uskottavuusfunktioden kohdalla y-akselien skaalat ovat tyystin erilaiset.



Kunkin yksittäisen satunnaismuuttujan Y_i tiheysfunktio on

$$g(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left(-\frac{1}{2\,\sigma^2}(y-\mu)^2\right).$$

Tässä exp tarkoittaa eksponenttifunktiota, eli

$$\exp(x) = \mathrm{e}^x, \qquad \text{kun } x \in \mathbb{R}.$$

Parametrien μ ja σ^2 merkitys on se, että kullakin i

$$EY_i = \mu, \qquad \operatorname{var} Y_i = \sigma^2.$$

Parametri μ on paitsi normaalijakauman $N(\mu,\sigma^2)$ odotusarvo, myös sen moodi ja mediaani. Normaalijakauman tiheysfunktio on symmetrinen odotusarvon suhteen. Varianssiparametri kuvaa sitä, miten tiukasti jakauma on keskittynyt keskikohtansa ympärille: mitä pienempi varianssi, sitä keskittyneempi jakauma.

Havaintosatunnaisvektorin $\mathbf{Y}$ yhteistiheysfunktio on

$$\begin{aligned} f(\mathbf{y};\mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left(-\frac{1}{2\,\sigma^2}(y_i-\mu)^2\right). \\ &= (2\,\pi\,\sigma^2)^{-n/2}\,\exp\left(-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \end{aligned} \tag{3.18}$$

Johdossa sovellettiin tuttua kaavaa

$$\mathrm{e}^a\,\mathrm{e}^b = \mathrm{e}^{a+b}, \quad \text{eli} \quad \exp(a)\,\exp(b) = \exp(a+b),$$

joka pätee kaikille reaaliluvuille a ja b.

Kaavasta (3.18) saadaan havaintoa $\mathbf{y}$ vastaavalle logaritmiselle uskottavuusfunktiolle lauseke

$$\begin{aligned} \ell(\mu,\sigma^2) &= \log f(\mathbf{y};\mu,\sigma^2) \\ &= -\frac{n}{2}(\log(2\,\pi) + \log(\sigma^2)) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \end{aligned} \tag{3.19}$$

Ylläolevassa kaavassa voidaan neliöiden summa hajottaa kahteen osaan (harjoitustehtävä)

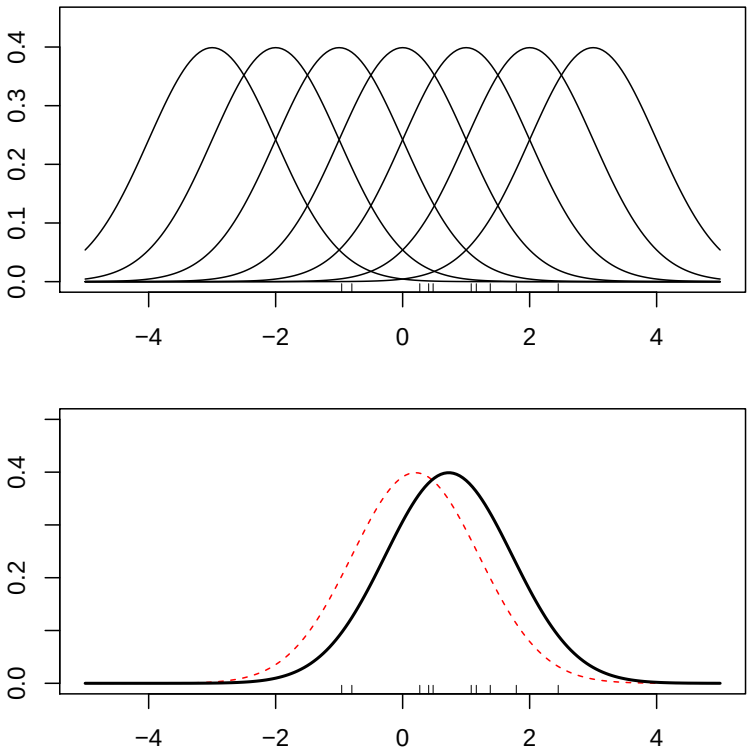
$$\sum_{i=1}^{n}(y_i-\mu)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 + n\,(\bar{y}-\mu)^2, \tag{3.20}$$

jossa $\bar{y}$ on lukujen y_i otoskeskiarvo,

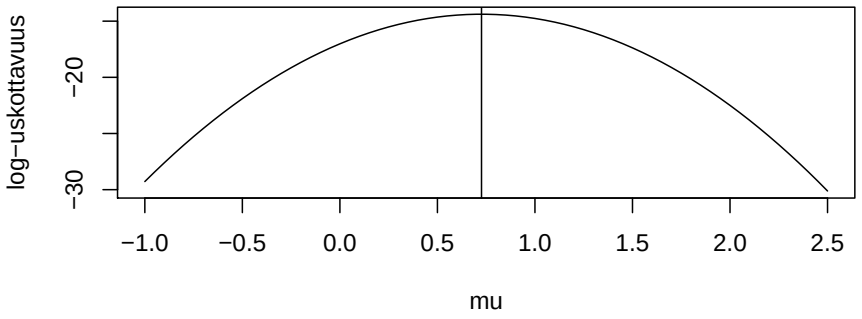
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.21}$$

Tämä huomio helpottaa SU-estimaattien löytämistä.

Kuva 3.3 Parametrin μ estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun σ^2 on tunnettu luku (tässä $\sigma^2 = 1$). Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, 1)$ tiheysfunktioita muutamilla eri parametrin μ arvoilla sekä eräästä normaalijakaumasta $N(\mu, 1)$ simuloitu aineisto (lyhet viivat x-akselin yläpuolella). Alemmassa kuvassa on paljastettu todellinen simuloinnissa käytetty tiheysfunktio (katkoviiva) sekä SU-estimaattia vastaava estimoitu tiheysfunktio (yhtenäinen viiva). Todellisessa tilastollisen päättelyn tilanteessa katkoviivalla merkittyä todellista tiheysfunktiota ei tunnettaisi.



Kuva 3.4 Parametrin μ logaritminen uskottavuusfunktio. SU-estimaatti on merkitty pystyviivalla.



3.8.1 Varianssi tunnettu

Jos normaalijakaumaperheessä varianssi σ^2 on tunnettu luku, niin mallissa on jäljellä vain yksi tuntematon parametri μ. Kuvassa 3.3 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää nyt valita kuvaamaan x-akselille lyhyillä viivoilla merkittyä aineistoa.

Logaritminen uskottavuusfunktio on kaavojen (3.19) ja (3.20) mukaan

$$\begin{aligned}\ell(\mu) &= -\frac{n}{2}(\log(2\,\pi) + \log(\sigma^2)) - \frac{1}{2\,\sigma^2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + n\,(\bar{y} - \mu)^2\right) \\ &= \text{vakio} - \frac{n}{2\,\sigma^2}(\bar{y} - \mu)^2\end{aligned}$$

Koska kerroin $n/(2\,\sigma^2)$ on positiivinen, niin logaritminen uskottavuusfunktio maksimoituu täsmälleen silloin, kun lauseke $(\bar{y} - \mu)^2$ minimoituu, eli silloin, kun $\mu = \bar{y}$. Logaritminen uskottavuusfunktio on esitetty kuvassa 3.4 kuvan 3.3 aineistolle.

Tässä tapauksessa SU-estimaatti on *otoskeskiarvo* (engl. *sample mean; average*), eli

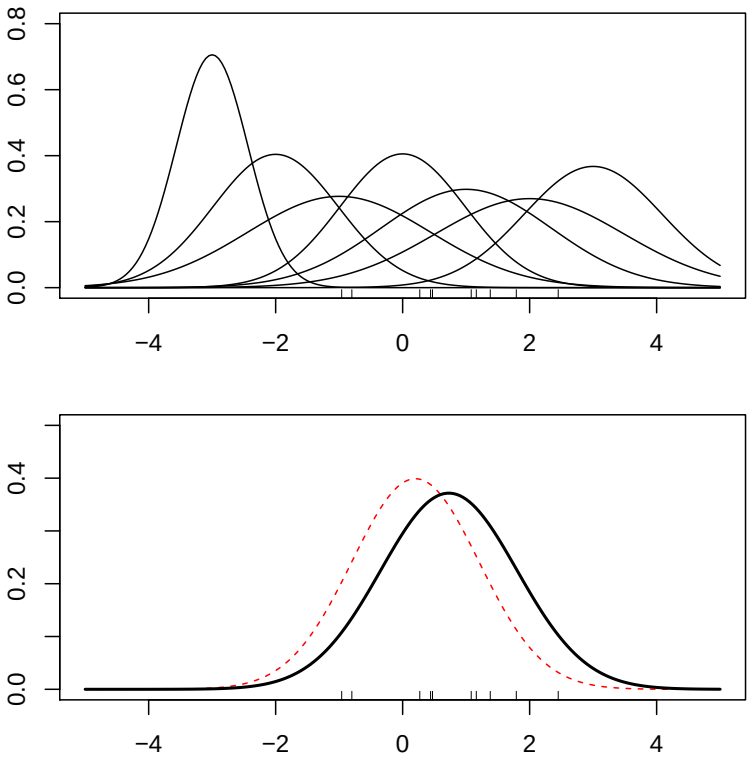
$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{3.22}$$

Vastaava estimaattori $\frac{1}{n}\sum_{i=1}^{n} Y_i$ on harhaton, ja sen varianssi on

$$\text{var}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sigma^2,$$

joka on tässä mallissa tunnettu vakio. Tämän luvun neliöjuuri on SU-estimaatin keskivirhe.

Kuva 3.5 Parametrin (μ, σ^2) estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun sekä μ että σ^2 ovat tuntemattomia. Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, \sigma^2)$ tiheysfunktioita muutamilla eri parametrivektorin (μ, σ^2) arvoilla sekä eräästä normaalijakaumasta simuloitu aineisto (lyhet viivat x-akselin yläpuolella). Alemmassa kuvassa on paljastettu todellinen simuloinnissa käytetty tiheysfunktio (katkoviiva) sekä SU-estimaattia vastaava estimoitu tiheysfunktio (yhtenäinen viiva). Todellisessa tilastollisen päättelyn tilanteessa katkoviivalla merkittyä todellista tiheysfunktiota ei tunnettaisi.



3.8.2 Molemmat parametrit tuntemattomia

Nyt molemmat parametri μ ja σ^2 ovat tuntemattomia, joten satunnaisvektorin $\mathbf{Y}$ jakauman kiinnittämiseksi pitäisi tuntea parametrivektorin $\boldsymbol{\theta} = (\mu, \sigma^2)$ arvo. Kuvassa 3.5 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää jälleen valita kuvaamaan x-akselille lyhyillä viivoilla merkittyä aineistoa.

Logaritminen uskottavuusfunktio on kaavojen (3.19) ja (3.20) mukaan

$$\ell(\mu, \sigma^2) = -\frac{n}{2}(\log(2\,\pi) + \log(\sigma^2)) - \frac{1}{2\,\sigma^2}\,\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{n}{2\,\sigma^2}\,(\bar{y} - \mu)^2$$

Tästä kaavasta nähdään, että vektori

$$\left(\bar{y},\ \sum_{i=1}^{n}(y_i - \bar{y})^2\right)$$

on tyhjentävä tunnusluku. Logaritminen uskottavuusfunktio on esitetty kuvassa 3.6 kuvan 3.3 aineistolle.

Logaritminen uskottavuusfunktio riippuu μ:n arvosta vain sen viimeisen termin kautta. Oli varianssiparametrin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion $\mu \mapsto \ell(\mu, \sigma^2)$ maksimoi arvo $\hat{\mu} = \bar{y}$. Tämän ansiosta maksimointi saadaan palautettua yhdestä muuttujasta riippuvan funktion u maksimointitehtäväksi, jossa

$$\begin{aligned} u(\sigma^2) &= \max_{\mu} \ell(\mu, \sigma^2) = \ell(\bar{y}, \sigma^2) \\ &= -\frac{n}{2}(\log(2\,\pi) + \log(\sigma^2)) - \frac{1}{2\,\sigma^2}\,\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned}$$

Tämän funktion maksimi puolestaan löytyy pisteestä

$$\hat{\sigma}^2 = \frac{1}{n}\,\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Näiden tarkastelujen jälkeen ollaan saatu selville, että parametrin (μ, σ^2) SU-estimaatti on

$$\hat{\mu} = \bar{y} = \frac{1}{n}\,\sum_{i=1}^{n} y_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\,\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{3.23}$$

Estimaattori

$$\hat{\mu}(\mathbf{Y}) = \frac{1}{n}\,\sum_{i=1}^{n} Y_i = \bar{Y}$$

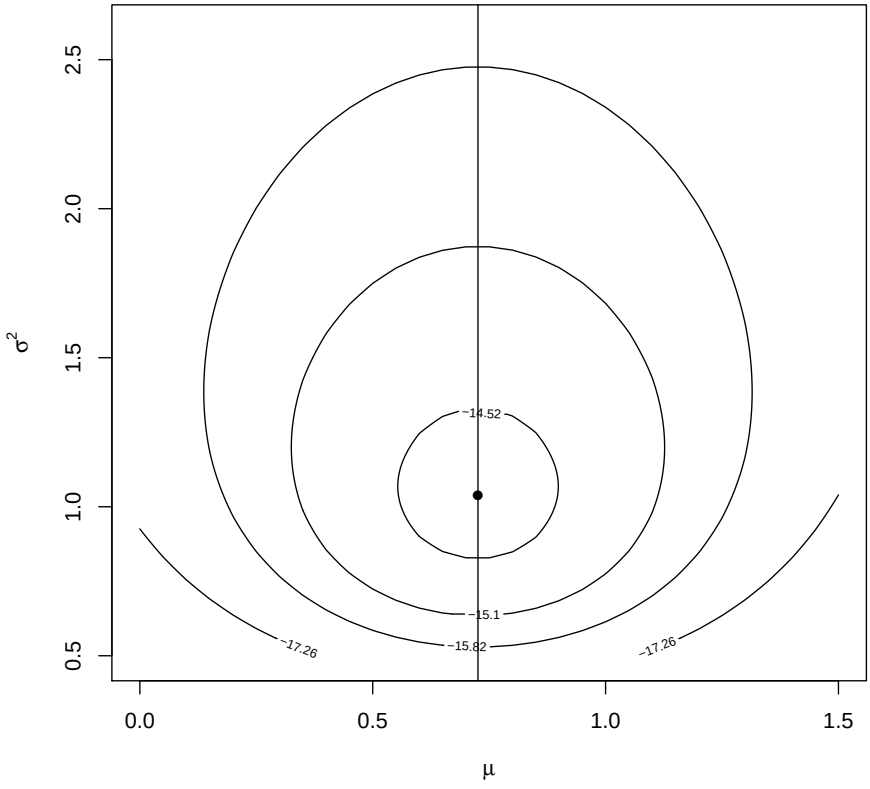
on harhaton, mutta varianssiparametrin SU-estimaattori

$$\hat{\sigma}^2(\mathbf{Y}) = \frac{1}{n}\,\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

on harhainen, sillä sen odotusarvo on

$$E_{(\mu,\sigma^2)}[\hat{\sigma}^2(\mathbf{Y})] = \frac{n-1}{n}\sigma^2.$$

Kuva 3.6 Parametrivektorin (μ, σ^2) logaritminen uskottavuusfunktio $\ell(\mu, \sigma^2)$ esitettynä tasa-arvokäyriensä avulla. SU-piste on merkitty pallolla. Millä tahansa varianssiparametrin arvolla funktion $\mu \mapsto \ell(\mu, \sigma^2)$ maksimi löytyy pisteestä $\mu = \bar{y}$, joka on osoitettu suoralla.



Koska harhan saa helposti korjattua, niin varianssin estimaattina käytetään tavallisesti SU-estimaatin sijasta *otosvarianssia* (engl. *sample variance*)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{3.24}$$

Sitä vastaava estimaattori

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{3.25}$$

on harhaton (varianssiparametrille σ^2), sillä

$$E_{(\mu,\sigma^2)}[S^2] = E_{(\mu,\sigma^2)}[\frac{n}{n-1}\hat{\sigma}^2(\mathbf{Y})] = \frac{n}{n-1}\,\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Estimaattorien $(\bar{Y}, S^2)$ yhteisotantajakauma tunnetaan. Esim. aineopintojen todennäköisyyslaskennan kurssilla todistetaan, että kun (mallin oletusten mukaan) $Y_1, \ldots, Y_n$ ovat riippumattomia normaalijakaumaa $N(\mu, \sigma^2)$ noudattavaa satunnaismuuttujaa, niin tällöin

$$\bar{Y} \text{ ja } S^2 \text{ ovat riippumattomia,} \tag{3.26}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{1}{n}\,\sigma^2), \tag{3.27}$$

$$\frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}. \tag{3.28}$$

Tässä χ^2_{n-1} tarkoittaa khiin neliön jakaumaa vapausasteluvulla $n-1$, joka on eräs kuuluisa positiivisella reaaliakselilla määritelty jatkuva jakauma. Sovellamme näitä tietoja myöhemmin.

Keskiarvoa $\bar{Y}$ koskeva jakaumatulos (3.27) on helppo johtaa. Normaalijakauman yhteenlaskuominaisuuden mukaan riippumattomien satunnaismuuttujien Y_1 ja Y_2 summalla on normaalijakauma, jonka parametrit saadaan laskemalla yhteen Y_1:n ja Y_2:n jakaumien parametrit, eli

$$Y_1 + Y_2 \sim N(\mu + \mu, \sigma^2 + \sigma^2).$$

(Varoitus: tämä on nimenomaan normaalijakaumaa, riippumattomia satunnaismuuttujia ja yhteenlaskua koskeva ominaisuus. Vastaavat kaavat eivät automaattisesti pidä paikkaansa muille jakaumille, riippuville satunnaismuuttujille, tai muille laskutoimituksille.) Tätä päättelyä voidaan jatkaa, jolloin summan jakaumaksi saadaan

$$Y_1 + \cdots + Y_n \sim N(n\mu, n\sigma^2).$$

Kun nyt muistetaan, että tässä ensimmäinen parametri on odotusarvo ja toinen varianssi, niin nähdään helposti, että luvulla $1/n$ skaalatun summan jakauma on

$$\bar{Y} \sim N(\mu, \frac{1}{n}\,\sigma^2).$$

Sen sijaan satunnaismuuttujan S^2 jakauman johtaminen on paljon monimutkaisempaa, ja se väite, että $\bar{Y}$ ja S^2 ovat riippumattomia voi ensinäkemältä herättää hämmennystä, sillä satunnaismuuttuja S^2 määritellään satunnaismuuttujan $\bar{Y}$ avulla.

Usein normaalijakaumamallissa ollaan tosiasiassa kiinnostuneita lähinnä populaation odotusarvosta μ, ja populaation varianssi σ^2 on ns. *haittaparametri* (engl. *nuisance parameter*), joka tarvitaan mallin spesifioimiseksi, mutta jonka arvosta ei olla kiinnostuneita. Tässä tapauksessa parametrin μ estimaatti on otoskeskiarvo $\bar{y}$. Vastaavan estimaattorin $\bar{Y}$ (otantajakauman) varianssi on σ^2/n. Kun tähän kaavaan sijoitetaan tuntemattoman populaatiovarianssin σ^2 tilalle sen otosestimaatti s^2, päädytään siihen, että keskiarvon keskivirhe lasketaan kaavalla

$$\frac{1}{\sqrt{n}}\, s,$$

jossa *otoskeskihajonta* (engl. *sample standard deviation*) s on otosvarianssin (3.24) neliöjuuri, eli

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3.29}$$

Otoskeskihajonta estimoi populaation keskihajontaa. Sen sijaan *keskiarvon keskivirhe* (engl. *standard error of the mean*)

$$\frac{1}{\sqrt{n}}\, s = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.30}$$

estimoi satunnaismuuttujan $\bar{Y}$ keskihajontaa $\sigma/\sqrt{n}$.

Jos odotusarvo on tuntematon, niin myös muulloin kuin normaalijakautuneen populaation tapauksessa populaation varianssia usein estimoidaan otosvarianssilla s^2 (3.24), jota vastaava estimaattori S^2 (3.25) on populaation varianssin harhaton estimaattori aina, kun käsitellään satunnaisotosta populaatiosta, jonka varianssi on σ^2. Populaation keskihajontaa $\sigma = \sqrt{\sigma^2}$ on myös tapana estimoida otoskeskihajonnalla, vaikka vastaava estimaattori $S = \sqrt{S^2}$ ei ole harhaton.

3.9 SU-estimaatteja muissa tilanteissa

3.9.1 Satunnaisotos eksponenttijakaumasta

Eksponenttijakaumaa noudattava satunnaismuuttuja X voi saada kaikkia positiivisia reaaliarvoja, ja sillä on tiheysfunktio

$$f(x;\lambda) = \lambda\, \mathrm{e}^{-\lambda\, x}, \qquad x > 0. \tag{3.31}$$

jossa jakauman parametria on merkitty kirjaimella $\lambda > 0$. Jakaumasta käytetään lyhennettä $\mathrm{Exp}(\lambda)$. Jos $X \sim \mathrm{Exp}(\lambda)$, niin sen odotusarvo ja varianssi ovat

$$EX = \frac{1}{\lambda}, \qquad \mathrm{var}\, X = \frac{1}{\lambda^2} = (EX)^2. \tag{3.32}$$

Olkoon $Y_1, \ldots, Y_n$ on satunnaisotos eksponenttijakaumasta $\mathrm{Exp}(1/\theta)$, jonka odotusarvo on θ on valittu mallin parametriksi. Tällöin SU-estimaattoriksi saadaan helpoilla laskuilla

$$\hat{\theta}(\mathbf{Y}) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.33}$$

SU-estimaattorin otantajakauma saadaan selvitettyä gammajakauman ominaisuuksien avulla. Gammajakauman Gamma(α, λ) (jossa $\alpha > 0, \lambda > 0$) tiheysfunktio on

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, \mathrm{e}^{-\lambda x}, \quad x > 0. \tag{3.34}$$

Jos $X \sim \text{Gamma}(\alpha, \lambda)$, niin sen odotusarvo ja varianssi ovat

$$EX = \frac{\alpha}{\lambda}, \qquad \text{var}\, X = \frac{\alpha}{\lambda^2}. \tag{3.35}$$

Tiheysfunktioiden kaavoja vertaamalla nähdään, että Exp(λ) on sama jakauma kuin Gamma$(1, \lambda)$. Aiemmin mainittu khiin neliön jakauma χ^2_ν vapausasteluvulla $\nu > 0$ on tietty gammajakauma, nimittäin

$$\chi^2_\nu = \text{Gamma}(\frac{1}{2}\nu, \frac{1}{2}). \tag{3.36}$$

Gammajakaumalla on seuraava yhteenlaskuominaisuus. Jos X_1 ja X_2 ovat riippumattomia gammajakautuneita satunnaismuuttujia, joilla jälkimmäinen parametri on sama, eli jos

$$X_1 \sim \text{Gamma}(\alpha_1, \lambda), \quad X_2 \sim \text{Gamma}(\alpha_2, \lambda), \quad X_1 \perp\!\!\!\perp X_2,$$

niin tällöin

$$X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda).$$

Toinen mukava ominaisuus on seuraava skaalausominaisuus. Jos $X \sim \text{Gamma}(\alpha, \lambda)$, ja $k > 0$ on vakio, niin

$$k\, X \sim \text{Gamma}(\alpha, \lambda/k).$$

Näiden ominaisuuksien avulla on helppo selvittää, että kun käsitellään satunnaisotosta eksponenttijakaumasta Exp$(1/\theta)$, niin SU-estimaattorin otantajakauma on

$$\bar{Y} \sim \text{Gamma}(n, n/\theta). \tag{3.37}$$

Estimaattori on harhaton, ja sen varianssi on $\frac{1}{n}\theta^2$. Estimaatin keskivirheen voi laskea joko kaavalla

$$\frac{1}{\sqrt{n}}\, \hat{\theta}$$

tai käyttämällä populaatiovarianssin estimaattina otosvarianssia, jolloin keskivirheelle saadaan kaava

$$\frac{1}{\sqrt{n}}\, s,$$

jossa s on otoskeskihajonta.

3.9.2 Satunnaisotos Poissonin jakaumasta

Poissonin jakaumaa noudattava satunnaismuuttuja X voi saada minkä tahansa kokonaislukuarvon $0, 1, 2, \ldots$, ja sillä on pistetodennäköisyysfunktio

$$f(x; \mu) = \mathrm{e}^{-\mu}\, \frac{\mu^x}{x!}, \qquad x = 0, 1, 2, \ldots, \tag{3.38}$$

jossa jakauman parametria on merkitty kirjaimella $\mu > 0$. Jakaumasta käytetään lyhennettä Poisson(μ). Jos $X \sim \text{Poisson}(\mu)$, niin

$$EX = \mu, \qquad \text{var}\, X = \mu = EX. \tag{3.39}$$

Jos $Y_1, \ldots, Y_n$ on satunnaisotos jakaumasta Poisson(θ), $\theta > 0$, niin tällöin SU-estimaattoriksi saadaan helpoilla laskuilla

$$\hat{\theta}(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{3.40}$$

Tällä estimaattorilla on diskreetti otantajakauma.

SU-estimaattorin otantajakauma saadaan selvitettyä Poissonin jakauman ominaisuuksien avulla. Poissonin jakaumalla on nimittäin seuraava yhteenlaskuominaisuus. Jos

$$X_1 \sim \text{Poisson}(\mu_1), \quad X_2 \sim \text{Poisson}(\mu_2), \quad X_1 \perp\!\!\!\perp X_2,$$

niin tällöin

$$X_1 + X_2 \sim \text{Poisson}(\mu_1 + \mu_2)$$

Kun $Y_1, \ldots, Y_n$ on satunnaisotos jakaumasta Poisson(θ), niin tällöin Poissonin jakauman yhteenlaskuominaisuuden nojalla

$$\sum_{i=1}^{n} Y_i \sim \text{Poisson}(n\,\theta),$$

joten SU-estimaattorin otantajakauman ptnf on

$$P_\theta \left[\bar{Y} = \frac{k}{n} \right] = \exp(-n\,\theta)\, \frac{(n\,\theta)^k}{k!}, \qquad k = 0, 1, 2, \ldots \tag{3.41}$$

SU-estimaattori on harhaton, ja sen varianssi on $\frac{1}{n}\theta$. Keskivirhe voidaan laskea jommalla kummalla seuraavista lausekkeista

$$\frac{1}{\sqrt{n}} \sqrt{\hat{\theta}}, \qquad \frac{1}{\sqrt{n}}\, s.$$

3.9.3 Huomautus

Lukijalle saattaa näistä esimerkeistä syntyä sellainen kuva, että SU-estimaatit saadaan laskettua joka tilanteessa jollakin yksinkertaisella kaavalla, ja että SU-estimaattorin otantajakauma tunnetaan aina. Tämä on harhaluulo. Yksinkertaisia kaavoja SU-estimaateille tunnetaan vain harvoissa tilanteissa. Tällöinkään ei aina tunneta SU-estimaattorin otantajakaumaa. Monimutkaisissa parametrisissa malleissa SU-estimaatit joudutaan yleensä hakemaan tietokoneella käyttämällä numeerisia optimointimenetelmiä. Samalla on mahdollista laskea approksimaatio estimaatin keskivirheelle.

3.10 Momenttimenetelmä

Momenttimenetelmä (engl. *method of moments*) on SU-menetelmää varhaisempi menetelmä estimaattorin määrittelemiseksi. Tarkastelemme tätä menetelmää siinä tapauksessa, jossa käsitellään satunnaisotosta $Y_1, \ldots, Y_n$ jakaumasta, jonka ptnf/tf on $g(y; \boldsymbol{\theta})$. Otamme käyttöön vielä satunnaismuuttujan Y jolla myöskin on on ptnf/tf $g(y; \boldsymbol{\theta})$.

Populaation k:s momentti ($k = 1, 2, \ldots$) määritellään kaavalla

$$\mu_k(\boldsymbol{\theta}) = EY^k = \begin{cases} \sum_y y^k \, g(y; \boldsymbol{\theta}) & \text{jos jakauma on diskreetti,} \\ \int y^k \, g(y; \boldsymbol{\theta}) \, \mathrm{d}y & \text{jos jakauma on jatkuva.} \end{cases} \tag{3.42}$$

Momenttia $\mu_k(\theta)$ voidaan estimoida k:nnella otosmomentilla

$$m_k(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^k, \tag{3.43}$$

joka on populaatiomomentin $\mu_k(\boldsymbol{\theta})$ harhaton estimaattori.

Momenttimenetelmässä estimaatti (tai estimaattori) muodostetaan ratkaisemalla yhtälöryhmästä

$$\begin{cases} \mu_1(\boldsymbol{\theta}) &= m_1 \\ \mu_2(\boldsymbol{\theta}) &= m_2 \\ & \vdots \\ \mu_r(\boldsymbol{\theta}) &= m_r \end{cases} \tag{3.44}$$

tuntematon suure $\boldsymbol{\theta}$, jossa otosmomentit $m_1, \ldots, m_r$ lasketaan aineistosta. Ehtoja asetetaan niin monta, että yhtälöryhmällä on yksikäsitteinen ratkaisu parametriavaruudessa. Tavallisesti yhtälöitä asetetaan niin monta, kuin parametrivektorissa on komponentteja.

Tällä tavalla saadaan aikaan näppäriä kaavoja estimaateille joissakin sellaisissa tilanteissa, joissa SU-estimaatit jouduttaisiin määrittämään numeerisesti.