Johdatus tilastolliseen päättelyyn

Petri Koistinen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

13. maaliskuuta 2013

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 arvoida, toteutuvatko kyseisen menetelmän oletukset. Mikäli oletukset eivät täyty, saattaa tilanteeseen sopivan menetelmän rakentelu vaatia pitkän tutkimushankkeen. Sitä paitsi tilastollista päättelyä voidaan lähestyä ainakin kahdesta aivan erilaisesta lähtökohdasta, joista keittokirjamaisissa oppikirjoissa tavallisesti esitetään vain yksi.

Nämä kaksi pääasiallista lähestymistapaa tilastolliseen päättelyyn ovat *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 peruskäsitteitä ja perusmenetelmiä, 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. Kirjoittaessani olen myös pitänyt käsillä seuraavia englanninkielisiä oppikirjoja: Arnold [1], Casella ja Berger [2], Davison [3], Kalbfleisch [4], Ross [6], Vidakovic [9]. Todennäköisyyslaskennan osalta oletan lukijan osapuilleen ymmärtävän P. Tuomisen kirjan Todennäköisyyslaskenta I [8] sisällön.

Sellaiselle lukijalle, joka tahtoo tutustua tilastollisen päättelyn kannalta tärkeisiin henkilöihin, suosittelen yleistajuisia kirjoja [7] ja [5].

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] J. G. Kalbfleisch. *Probability and Statistical Inference II.* Springer, 1979.

[5] Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.* Yale University Press, 2011.

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

[7] David Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* W. H. Freeman, 2001.

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

[9] Brani Vidakovic. *Statistics for Bioengineering Sciences with MATLAB and WinBUGS Support.* Springer Texts in Statistics. Springer, 2011.

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).

> Tilastollisen päättelyn perusajatus on se, että 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 $y_1, \ldots, y_n$. Edellä $\mathbf{y} = (y_1, \ldots, y_n)$ 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 parametrin θ arvo 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 todennäköisyysmallin $f(\mathbf{y}; \theta)$ ajatellaan useimmiten 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 olettaa, 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 olettaa, 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 $0, 1, \ldots, N$.

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).

Yhteisjakauma voitaisiin 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, ts.

$$\theta = P(Y_i = 1).$$

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 (engl. *binomial experiment*):

- 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 ($Y_i = 0$). (Tällaista koetta kutsutaan Bernoullin kokeeksi.)
- Peräkkäisten toistokokeiden tulokset oletetaan toistaan riippumattomiksi, kun koetta kuvaava parametrin arvo on kiinnitetty.

Jos p on onnistumistodennäköisyys yhdessä toistossa, niin satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteisjakaumalla on yptf

$$\begin{aligned} f(\mathbf{y}; p) &= p^{y_1} (1-p)^{1-y_1} \cdots p^{y_n} (1-p)^{1-y_n} \\ &= p^{t(\mathbf{y})} (1-p)^{n-t(\mathbf{y})}, \end{aligned} \tag{2.6}$$

jossa

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

on onnistumisten lukumäärä (ykkösten lukumäärä) vektorissa $\mathbf{y}$. Pallot kulhossa -esimerkissä $p = \theta/N$, mutta nasta purkissa -esimerkissä oli $p = \theta$.

Binomikokeessa 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 $0 \leq p \leq 1$ on onnistumistodennäköisyys (ykkösen todennäköisyys) yhdessä kokeessa. Lyhyemmin merkittynä

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

Binomijakauman pistetodennäköisyysfunktio on

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{2.7}$$

Tästä näkökulmasta ainoa oleellinen ero edellisten jaksojen esimerkkien välillä on se, että pallot kulhossa -esimerkissä parametriavaruus on diskreetti, mutta nasta purkissa -esimerkissä parametriavaruus on jatkuva.

2.5 Kaksi lähestymistapaa päättelyyn

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

Estimointiteoriaa

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 tarkoituksella lainausmerkkeihin. 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 (1919–) 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

$$EY_i = p, \quad \operatorname{var} Y_i = p(1-p), \quad \text{jossa } p = \frac{\theta}{N}.$$

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

$$EX = np, \quad \operatorname{var} X = np(1-p), \quad \text{jossa } p = \frac{\theta}{N}.$$

Kaikkia näitä suureita voidaan kutsua parametreiksi. Parametri on yleisesti ottaen jokin mallin parametrista θ riippuva lauseke $\tau = k(\theta)$. Parametreja merkitään usein 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 havaintojen avulla. Tällöin soveltajan tulee tarkoin miettiä, mitä populaatiota havaintoaineisto edustaa, eli mihin populaatioon tilastolliset johtopäätökset voidaan yleistää.

Määritelmä 3.1 (Tunnusluku)**.** Tunnusluku (engl. *statistic*) tarkoittaa mitä tahansa lukua, joka voidaan laskea aineistosta ilman, että tarvitsee tuntea mitään tilastollisen mallin tuntematonta parametria.

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 3.2 (Estimaatti)**.** Joitakin tunnuslukuja käytetään parametrien arvioina, jolloin niitä kutsutaan vastaavien parametrien estimaateiksi (engl. *estimate*).

Esimerkki 3.1 (Onnistumistodennäköisyyden estimointi nasta purkissa -esimerkissä)

- Onnistumistodennäköisyyttä θ binomikokeessa arvioidaan tavallisesti laskemalla onnistumisten suhteellinen osuus n kokeessa, eli

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

- Esimerkiksi luku θ ei ole parametrin θ estimaatti, sillä θ ei ole tunnusluku: sitä ei voida laskea aineistosta.

△

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 N tarkoitti pallojen kokonaislukumäärää, n nostojen lukumäärää ja θ valkoisten pallojen lukumäärää. Jos onnistumistodennäköisyyttä $\phi = \theta/N$ estimoidaan onnistumisten suhteellisella osuudella (3.1), niin tällöin törmätään siihen ongelmaan, että tämän parametrin ϕ arvot kuuluvat 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 kahdella joukolla välttämättä ole edes kovin montaa yhteistä alkiota. Tämä ongelma voitaisiin 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 (joskus on mielekästä laajentaa tarkastelu mallin ulkopuolelle), oletetaan että satunnaisvektorilla $\mathbf{Y}$ on todellista parametrinarvoa θ vastaava todennäköisyysjakauma.

Määritelmä 3.3 (Otantajakauma)**.** Satunnaismuuttujan $t(\mathbf{Y})$ jakaumaa kutsutaan tämän tunnusluvun otantajakaumaksi (engl. *sampling distribution*). Vastaavasti, muotoa $h(\mathbf{Y}, \theta)$ olevan suureen jakaumaa kutsutaan tämän suureen otantajakaumaksi. Oletamme, että $\mathbf{Y}$ noudattaa jakaumaa $f(\mathbf{y}; \theta)$ todellisella parametrinarvolla θ.

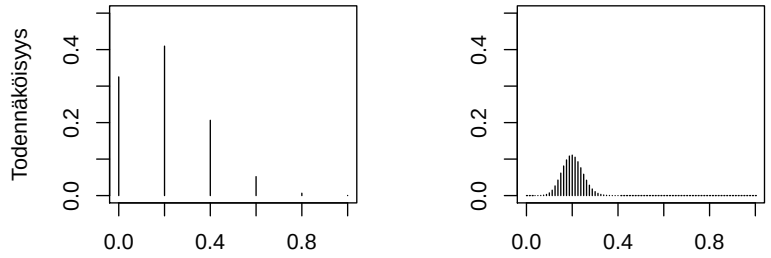
Termissä otantajakauma on taustalla ajatus otannan tai aineiston keruun toistamisesta. Jos aineiston keruu voitaisiin toistaa samoissa olosuhteissa riippumattomasti r kertaa, ja saataisiin aineistot $\mathbf{y}_1, \ldots, \mathbf{y}_r$ (jossa kukin $\mathbf{y}_i$ on n-vektori), niin tällöin arvot $t(\mathbf{y}_1), \ldots t(\mathbf{y}_r)$ 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 $f(\mathbf{y}; \theta)$.

Huomautuksia:

- Parametri θ on frekventistisessä päättelyssä kiinteä mutta tuntematon luku (tai vektori jonka komponentit ovat lukuja). Sillä ei ole todennäköisyysjakaumaa.

- Frekventistisen päättelyn teoriassa tarkastellaan parametriavaruudessa määriteltyjä jakauma. Ne ovat aina otantajakaumia: joko jonkin tunnusluvun $t(\mathbf{Y})$ tai jonkin suureen $h(\mathbf{Y}, \theta)$ otantajakaumia.

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



Frekventistisessä tilastotieteessä erityisen kiinnostava asia on estimaattorin otantajakauma. Sana estimaattori (engl. *estimator*) tarkoittaa sitä, että emme ajattele konkreettista lukua $\hat{\theta} = t(\mathbf{y})$ eli parametrin θ estimaattia, vaan tarkastelemme vastaavaa satunnaismuuttujaa $t(\mathbf{Y})$ eli estimaattoria.

Kirjallisuudessa merkintä $\hat{\theta}$ tarkoittaa toisinaan estimaattia ja toisinaan estimaattoria. Koska tässä vaiheessa vasta opettelemme käyttämään näitä käsitteitä, on hyödyllistä tehdä merkinnöissä ero näiden kahden asian välillä. Myöhemmässä vaiheessa on luvallista yksinkertaistaa merkintöjä. Tässä tekstissä $\hat{\theta}$ tai $\hat{\theta}(\mathbf{y})$ tarkoittaa estimaattia (ts. konkreettista lukua), ja $\hat{\theta}(\mathbf{Y})$ vastaavaa estimaattoria, joka on satunnaismuuttuja. Jos $\hat{\theta}$ lasketaan aineistosta $\mathbf{y}$ kaavalla $t(\mathbf{y})$, niin $\hat{\theta}(\mathbf{Y}) = t(\mathbf{Y})$.

Nasta purkissa -esimerkissä estimaattia

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

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

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

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

3.3 Todennäköisyyslaskennan kertausta

Jos X on satunnaismuuttuja, jonka odotusarvo on $\mu = EX$, niin X:n varianssi on luku

$$\mathrm{var}(X) = E(X-\mu)^2.$$

Jos X on satunnaismuuttuja, ja a ja b ovat vakiota, niin satunnaismuuttujan $a\,X + b$ odotusarvo ja varianssi ovat

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

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

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

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

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

Tšebyševin epäyhtälön mukaan mille tahansa vakiolle a

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

3.4 Otantajakauman ominaisuuksia

Binomikokeessa estimaattorin

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

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

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

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

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 $\operatorname{Bin}(n, \theta)$ odotusarvolle ja varianssille, sillä estimaattori $\hat{\theta}(Y)$ on yhtä kuin $\frac{1}{n}X$, jossa satunnaismuuttuja $X = \sum_1^n Y_i$ noudattaa binomijakaumaa $\operatorname{Bin}(n, \theta)$.

Määritelmä 3.4 (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 (engl. *mean unbiased*). Odotusarvon sijasta voisimme toki tarkastella muitakin otantajakauman keskikohtaa kuvailevia suureita, erityisesti mediaania. Voisimme määritellä samaan tapaan, mitä tarkoittaa mediaanin mielessä harhaton (engl. *median unbiased*) estimaattori. Jätämme tämän tarkennuksen kuitenkin tekemättä.

Määritelmä 3.5 (Harha)**.** Estimaattorin $\hat{\theta}(\mathbf{Y})$ harha on

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

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

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

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 näyttävät raskailta, joten avaan seuraavaksi niiden merkitystä estimaattorin harhan määritelmän eli kaavan (3.8)

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

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 tavalla tämä asia oppikirjoissa tavallisesti esitetään.

Määritelmä 3.6 (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. Keskineliövirheen määritelmän voi myöhemmässä vaiheessa lyhentää muotoon

$$\mathrm{mse}(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

Mikäli estimaattori on harhaton, niin sen keskineliövirhe on tietenkin sama asia kuin sen varianssi. Helpolla laskulla (vrt. kaava (3.6)) nähdään, että keskineliövirhe voidaan esittää laskemalla yhteen estimaattorin varianssi ja harhan neliö (engl. *bias-variance decomposition*), eli

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

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

Määritelmä 3.7 (Keskineliövirheen neliöjuuri, RMSE)**.** Estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirheen neliöjuuri (engl. *root mean squared error*) on

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

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

Määritelmä 3.8 (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.*

Monimutkaisissa tapauksissa todennäköisyyslaskennan taitomme eivät enää riitä estimaattorin otantajakauman selvittämiseen. Tällöin niitä voidaan yrittää selvittää tietokonesimuloinnin avulla. Toinen mahdollisuus on soveltaa asymptoottisia (ts. suuren otoskoon) approksimaatioita otantajakaumalle.

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 Tarkentuvuus

Tarkentuvuus on yksinkertainen esimerkki estimaattorin asymptoottisesta ominaisuudesta. Tällöin otoskoon kuvitellaan kasvavan rajatta, vaikka todellisuudessa havaintoja tietenkin on täsmälleen vain niin monta kuin mitä niitä on. Estimaattorien asymptoottisia ominaisuuksia tarkastellaan sen takia, että riittävän suurella otoskoolla asymptoottisten ominaisuuksien ajatellaan toteutuvan likimäärin. Saman tien on syytä tunnustaa, että teoreettisessa tilastotieteessä ei tyypillisesti pystytä kertomaan, milloin otoskoko on riittävän suuri jotta asymptotiikasta saatava arvio olisi käytännön kannalta riittävällä tarkkuudella voimassa. Tämä on taas kerran sellainen asia, jota on helpointa yrittää selvittää tietokonesimuloinneilla.

Kun puhutaan estimaattorien asymptoottisista ominaisuuksista, niin kutakin otoskokoa n vastaa yksi estimaattori ja oikeastaan tällöin tarkastellaan näiden estimaattorien muodostamaa jonoa. Yksinkertaisuuden vuoksi emme merkitse otoskokoa näkyviin estimaattorin yhteyteen. Käytämme tavanomaista puhetapaa, jossa ei tehdä eroa yksittäistä otoskokoa vastaavan estimaattorin ja eri otoskokoja vastaavien estimaattorien muodostaman estimaattorijonon välillä.

Parametrin θ estimaattori $\hat{\theta}(\mathbf{Y})$ on tarkentuva (engl. *consistent*), mikäli se suppenee kohti parametrin θ todellista arvoa otoskoon n kasvaessa rajatta. Tällöin tietenkin myös havaintosatunnaisvektorin

$$\mathbf{Y} = (Y_1, \ldots, Y_n)$$

pituus kasvaa rajatta. Tarkentuvuus on oikeastaan edellytys sille, että estimaattorin $\hat{\theta}(\mathbf{Y})$ voidaan sanoa estimoivan parametria θ. Jos otokokoa pystytään kasvattamaan rajatta, niin parametrin arvo saadaan rajalla selvitettyä tarkentuvan estimaattorin avulla.

Tämän kurssin puitteissa tarkentuvuuden yhteydessä vaaditaan ns. stokastinen suppeneminen, joka määritellään ensin satunnaismuuttujajonolle $X_1, X_2, \ldots$.

Määritelmä 3.9 (Stokastinen suppeneminen)**.** Jono satunnaismuuttujia $X_1, X_2, \ldots$ suppenee stokastisesti (engl. *converges in probability*) kohti vakiota a, mikäli

$$P(|X_n - a| > \epsilon) \to 0, \qquad \text{kaikilla } \epsilon > 0.$$

Tämä asia voidaan ilmaista merkinnällä

$$X_n \xrightarrow{\mathrm{P}} a.$$

Jos $X_n \xrightarrow{\mathrm{P}} a$ ja $\epsilon > 0$ on mielivaltaisen pieni luku, niin todennäköisyys, että X_n ei satu välille $[a-\epsilon, a+\epsilon]$ suppenee kohti nollaa n:n kasvaessa. Yhtäpitävästi voidaan sanoa, että todennäköisyys, että X_n sattuu välille $[a-\epsilon, a+\epsilon]$ suppenee kohti ykköstä, kun n kasvaa rajatta. Tämä tarkoittaa sitä, että satunnaismuuttujien X_n jakauma keskittyy yhtä tiukemmin luvun a läheisyyteen, kun n kasvaa.

Stokastisen suppenemisen voi usein todistaa seuraavan kriteerin avulla.

$$E(X_n - a)^2 \to 0 \quad \Rightarrow \quad X_n \xrightarrow{\mathrm{P}} 0. \tag{3.13}$$

Tämä seuraa Tšebyševin epäyhtälöstä (3.7). Jos nimittäin $\epsilon > 0$, niin

$$P(|X_n - a| > \epsilon) \leq \frac{E(X_n - a)^2}{\epsilon^2},$$

ja tämä yläraja suppenee oletuksen mukaan kohti nollaa n:n kasvaessa.

Määritelmä 3.10 (Tarkentuvuus)**.** Jos $\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta$ kaikilla $\theta \in \Theta$, kun otoskoko n ja sen mukana havaintosatunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ pituus kasvavat rajatta, niin tällöin sanotaan, että estimaattori(jono) $\hat{\theta}(\mathbf{Y})$ on *tarkentuva* (engl. *consistent*).

Tyypillisesti estimaattorin keskineliövirhe $\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y}))$ suppenee kohti nollaa, kun otoskoko n kasvaa rajatta. Tällöin siis kaikilla θ pätee

$$E_\theta \left(\hat{\theta}(\mathbf{Y}) - \theta\right)^2 \to 0, \qquad \text{kun } n \to \infty.$$

Tuloksen (3.13) mukaan $\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta$ kaikilla θ joten tällöin ko. estimaattori on tarkentuva.

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 kaikilla $0 \leq \theta \leq 1$, on estimaattori tarkentuva.