# Johdatus tilastolliseen päättelyyn

Petri Koistinen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

29. huhtikuuta 2013

# Sisältö

# 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 $\Omega$ 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 $\theta$, 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 $\theta$ 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 $\theta$ 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 $\theta$.

Usein käsittelemme tilannetta, jossa satunnaismuuttujat $Y_i$ ovat riippumattomia ja niillä on sama jakauma, kun parametrinarvo $\theta$ 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 $\theta$ arvo tunnetaan. Ongelma syntyy siitä, että $\theta$ on tuntematon. Parametrista tiedetään kuitenkin vähintään sen verran, että osataan sanoa, missä joukossa $\Theta \subset \mathbb{R}$ sen arvot voivat olla. Tällaista joukkoa $\Theta$ 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 $\theta$ 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 $\theta$ on valkoista, niin on luonnollista olettaa, että

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

Edellä merkittiin parametri $\theta$ selvyyden vuoksi näkyviin todennäköisyyden $P(\cdot)$ alaindeksiksi. Jotta edellä kirjoitetulla todennäköisyydellä olisi numeerinen arvo, täytyy luvun $N$ sekä valkoisten pallojen lukumäärän $\theta$ 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 $\theta$ arvo on jokin luvuista $0, 1, \ldots, N$.

On hyödyllistä huomata, että yptnf voidaan esittää (yhdistämällä termien $\theta$ 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 $\theta$ 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 $\theta$ 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 $\theta$: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 $\theta$, 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 $\theta$ olisi yhtä kuin selälleen päätyvien tulosten suhteellinen osuus äärettömän pitkässä koesarjassa. Millään äärellisen pitkällä koesarjalla $\theta$: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 $\theta$ arvo tunnetaan, mutta tilastollisessa päättelyssä $\theta$ on tuntematon luku. Tämän takia ensimmäisenä pyrkimyksenä on arvioida eli estimoida parametrin $\theta$ 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 $\theta$ 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 $\Theta$. 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 $\theta$. 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) $\theta$, 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 $\theta$ 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 $\theta$ 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 $\theta$ ei ole parametrin $\theta$ estimaatti, sillä $\theta$ 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 $\theta$ estimaatin $\hat{\theta}$ pitää kuulua parametriavaruuteen $\Theta$. 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 $\theta$ valkoisten pallojen lukumäärää. Jos onnistumistodennäköisyyttä $\phi = \theta/N$ estimoidaan onnistumisten suhteellisella osuudella (3.1), niin tällöin törmätään siihen ongelmaan, että tämän parametrin $\phi$ 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 $\theta$ 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 $\theta$.

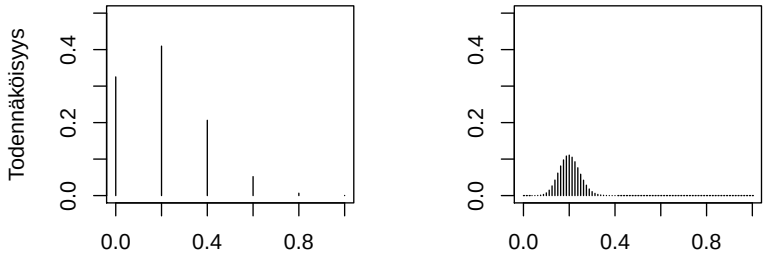
Termissä otantajakauma on taustalla ajatus otannan tai aineiston keruun toistamisesta. Jos aineiston keruu voitaisiin toistaa samoissa olosuhteissa riippumattomasti $r$ kertaa, ja saataisiin aineistot $\mathbf{y}_1, \ldots, \mathbf{y}_r$ (jossa kukin $\mathbf{y}_i$ on $n$-vektori), niin tällöin arvot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

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

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

**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 $\theta$ 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 $\theta$ 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 $\theta$, 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 $\theta$ estimaattori $\hat{\theta}(\mathbf{Y})$ on tarkentuva (engl. *consistent*), mikäli se suppenee kohti parametrin $\theta$ 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 $\theta$. 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.

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

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 $\theta$ 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 $\theta$ 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.

# Luku 4

# Suurimman uskottavuuden menetelmä ja momenttimenetelmä

## 4.1 Uskottavuusfunktio

Palautetaan ensin mieleen, että funktio

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

eli lauseke $f(\mathbf{y}; \theta)$ ymmärrettynä argumentin $\mathbf{y}$ funktiona kiinteällä $\theta$ on satunnaisvektorin $\mathbf{Y}$ yhteistiheysfunktio tai yhteispistetodennäköisyysfunktio.

Kun aineisto $\mathbf{y}$ on havaittu, ja havaittua arvoa käytetään funktion $f(\mathbf{y}; \theta)$ ensimmäisenä argumenttina, niin tämä lauseke on enää argumentin $\theta$ funktio. 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 $\theta$ uskottavuusfunktio.

Huomaa, että uskottavuusfunktion yhteydessä $\theta$ 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.

Vaikka funktio

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

on ypntf tai ytf kaikilla $\theta$, niin huomaa, että uskottavuusfunktio on funktio

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

ja se ei ole yptnf eikä ytf.

**Esimerkki 4.1** (Uskottavuusfunktio binomikokeessa) 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 binomikokeen yptnf:n kaavan (2.6) mukaan

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

Tässä onnistumistodennäköisyys on $p = \theta/N$, joka on valkoisten pallojen suhteellinen osuus kulhossa.

Jos taas sama havainto $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$ saadaan nasta purkissa -esimerkissä, niin tällöin onnistumistodennäköisyys on $\theta$, ja uskottavuusfunktio on

$$L(\theta) = \theta^2 (1 - \theta)^5.$$

Tässä parametriavaruudeksi ja uskottavuusfunktion määrittelyjoukoksi voidaan valita joko suljettu väli $[0, 1]$ tai avoin väli $(0, 1)$. △

Laajennamme uskottavuusfunktion määritelmää sillä tavalla, että uskottavuusfunktioksi kelpuutetaan myös mikä tahansa muotoa

$$L(\theta) = k(\mathbf{y})\, f(\mathbf{y}; \theta) \tag{4.1}$$

oleva lauseke, jossa positiivinen vakio $k(\mathbf{y}) > 0$ saa riippua aineistosta $\mathbf{y}$, mutta ei saa riippua uskottavuusfunktion argumentista $\theta$. Edellä tehtiin aina valinta $k(\mathbf{y}) = 1$.

Tilastollisessa päättelyssä yleensä suositaan sellaisia menetelmiä, jotka perustuvat ainoastaan uskottavuusfunktion käyttöön ja joiden kannalta on saman tekevää, mitä verrannollisuuskerrointa $k = k(\mathbf{y}) > 0$ uskottavuusfunktion määritelmässä käytetään. Esimerkiksi uskottavuusfunktion maksimikohta pysyy samana vaikka kerrointa $k > 0$ muutetaan. Uskottavuusfunktion (tai siis minkä tahansa uskottavuusfunktion version) voidaan ajatella sisältävän kaiken aineistoon liittyvän informaation parametrin $\theta$ arvosta.

Usein uskottavuusfunktion sijasta tarkastellaan sen logaritmia.

**Määritelmä 4.1** (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. Koska uskottavuusfunkio on määrätty vain positiivista verrannollisuuskerrointa $k > 0$ vaille, niin tämän seurauksena logartiminen uskottavuusfunktio on määrätty vain vakiota $\log k$ vaille; funktioon $\ell(\theta)$ voidaan lisätä mikä tahansa vakio, jos tämä yksinkertaistaa kaavoja.

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.

## 4.2 Suurimman uskottavuuden estimaatti

Frekventistisessä tilastotieteessä parametria $\theta$ 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

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta). \tag{4.2}$$

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

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

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

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta). \tag{4.3}$$

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

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

Tuntuu järkevältä suosia sellaisia parametrin arvioita, joille havainnot ovat todennäköisiä eikä sellaisia, joille ne ovat epätodennäköisiä. Koska parametriavaruudesta joudutaan yksi piste estimaatiksi valitsemaan, niin miksipä ei valittaisi sitä pistettä, joka tekee havainnot mahdollisimman todennäköisiksi.

Jatkuvan yhteisjakauman tapauksessa SU-menetelmän motivointi on samantapainen, mutta monimutkaisempi. Oletamme, että havaintosatunnaisvektorin yhteisjakauma on jatkuva ja että satunnaismuuttujat $Y_i$ ovat riippumattomia, kuten kaavassa (2.2). SU-estimaatti on se parametriarvo, joka maksimoi yhteistiheysfunktion arvon laskettuna aineistolle $\mathbf{y}$. Koska yhteisjakauma on jatkuva, niin yhteispistetodennäköisyys $P_\theta(\mathbf{Y} = \mathbf{y}) = 0$, joten tätä tarkastelemalla emme saa aikaan järkevää kriteeriä. Sen sijaan tarkastelemme todennäköisyyttä, että kukin satunnaismuuttuja $Y_i$ saa arvonsa sellaiselta lyhyeltä väliltä $[a_i, b_i]$, joka sisältää havainnon $y_i$.

$$\begin{aligned} P_\theta(a_1 \le & Y_1 \le b_1, a_2 \le Y_2 \le b_2, \ldots, a_n \le Y_n \le b_n) \\ &= \int_{a_1}^{b_1} f_{Y_1}(t_1;\theta)\,\mathrm{d}t_1 \int_{a_2}^{b_2} f_{Y_2}(t_2;\theta)\,\mathrm{d}t_2 \ \ldots \int_{a_n}^{b_n} f_{Y_n}(t_n;\theta)\,\mathrm{d}t_n \\ &\approx (b_1 - a_1) f_{Y_1}(y_1;\theta)\ (b_2 - a_2) f_{Y_2}(y_2;\theta)\ \ldots\ (b_n - a_n) f_{Y_n}(y_n;\theta) \\ &= f(\mathbf{y};\theta) \prod_{i=1}^{n} (b_i - a_i) \end{aligned}$$

Tässä ensin vedottiin satunnaismuuttujien $Y_i$ riippumattomuteen, ja sen jälkeen kussakin integraalissa tehtiin seuraava approksimaatio. Lyhen välin $[a_i, b_i]$ yli laskettu integraali funktiosta $f_{Y_i}(t_i;\theta)$ on (integraalilaskennan väliarvolauseeseen nojaten) osapuilleen sama kuin sen suorakaiteen pinta-ala, jonka kanta on kyseisen välin pituus ja korkeus $f_{Y_i}(y_i;\theta)$. (Tiheysfunktiot $y_i \mapsto f_{Y_i}(y_i;\theta)$ oletetaan jatkuviksi.) Todennäköisyydeksi saatiin osapuilleen välien pituuksien tulo kertaa yhteistiheysfunktion arvo $f(\mathbf{y};\theta)$. Tämän tarkastelun jälkeen näemme, että SU-menetelmän motivaatio on myös jatkuvan yhteisjakauman tapauksessa se, että yritämme valita sellaisen parametriavaruuden pisteen, joka tekee havainnot mahdollisimman todennäköiseksi.

**Esimerkki 4.2** (Jatkoa esimerkille 4.1, pallot kulhossa) Valkoisten pallojen lukumäärä $\theta$ 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 takia 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*.

Pallot kulhossa -esimerkissä parametriavaruus on diskreetti. Koska se ei esimerkissä koostu kovin monesta pisteestä, pystymme laskemaan uskottavuusfunktion arvon jokaisessa parametriavaruuden pisteessä. Tämän jälkeen valitsemme sen pisteen, jossa suurin arvo saavutetaan.

Jos parametriavaruus on jatkuva, niin tälläinen menettely ei tule kyseeseen, vaan maksimoinnissa käytetään hyväksi derivaattaa. Tarkastelmme ensin yhden parametrin $\theta$ tapausta. Yksinkertaisissa tilanteissa SU-estimaatti saadaan ratkaistua algebrallisesti etsimällä logaritmisen uskottavuusfunktion derivaatan nollakohdat, eli ratkaisemalla ns. uskottavuusyhtälö

$$\ell'(\theta) = 0.$$

Tämä perustuu siihen, että mikäli (jatkuvasti derivoituva) yhden muuttujan funktio saavuttaa maksimin jossakin määrittelyjoukkonsa sisäpisteessä, niin kyseisessä pisteessä funktion derivaatta saa arvon nolla. Tämän jälkeen pitää funktion kriittisistä pisteistä eli derivaatan nollakohdista valita ne, jotka ovat maksimipisteitä. Tämä onnistuu joko tarkastelemalla derivaatan merkkikaaviota tai $\ell$:n toista derivaattaa (jos $\ell'(\theta_0) = 0$ ja $\ell''(\theta_0) < 0$, niin $\theta_0$ on maksimipiste). Lisäksi pitää kiinnittä huomiota (log-)uskottavuusfunktion käyttäytmiseen, kun lähestytään parametriavaruuden reunapisteitä. Tällä tavalla löydetään kaikki paikalliset maksimipisteet, ja lopulta niistä valitaan globaali maksimi, eli se piste, jossa $\ell$ saavuttaa suurimman arvonsa koko parametriavaruudessa. Näemme tästä menettelystä esimerkkejä seuraavissa jaksossa.

Jos parametreja on useita, niin kaikki $\ell$:n ensimmäisen kertaluvun osittaisderivaatat häviävät maksimipisteessä, joten tällöin uskottavuusyhtälö on yhtälöryhmä. Esim. kahden parametrin $\theta = (\mu, \phi)$ tapauksessa pitäisi etsiä ne pisteet, joissa molemmat yhtälöt

$$\frac{\partial}{\partial \mu}\ell(\mu, \phi) = 0, \qquad \frac{\partial}{\partial \phi}\ell(\mu, \phi) = 0$$

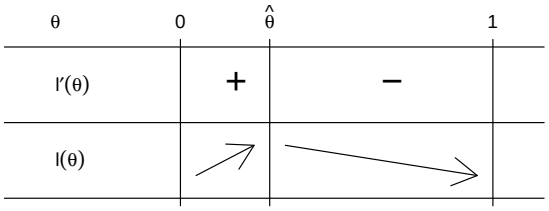
toteutuvat. Kriittisen pisteen laadun (minimi, maksimi, satulapiste) voi tarkistaa toisen kertaluvun osittaisderivaattojen avulla.

Monimutkaisemmissa tapauksissa maksimipisteitä ei enää pystytä määrittämään algebrallisesti, vaan ne haetaan tietokoneen avulla soveltamalla jotakin numeerista maksimointimenetelmää.

## 4.3 SU-estimaatti binomikokeessa

Binomikokeessa uskottavuusfunktion tai sen logaritmin arvo voidaan laskea kaavan (2.6) heti, kun tiedetään onnistumisten lukumäärä, toistojen lukumäärä

**Kuva 4.1** Logaritmisen uskottavuusfunktion kulkukaavio binomikokeessa, kun $n = 7$ toistossa havaitaan $x = 2$ onnistumista.




sekä parametriavaruus. Tätä varten ei tarvitse tietää, missä järjestyksessa onnistumiset ja epäonnistumiset sattuivat aineistossa $(y_1, \ldots, y_n)$. Johdamme seuraavaksi SU-estimaatin kaavan, kun $n$ toistossa onnistutaan $x$ kertaa, ja onnistumistodennäköisyys yhdessä toistossa on $\theta$. Oletamme, että parametriavaruus $\Theta$ on joko avoin väli $(0, 1)$ tai suljettu väli $[0, 1]$. Tällainen tilanne oli nasta purkissa -esimerkissä (mutta pallot pallot kulhossa -esimerkissä parametriavaruus oli diskreetti).

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

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

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

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

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

Derivaatan ainoa nollakohta on $\hat{\theta} = x/n$, ja kyseessä on maksimipiste, sillä derivaatan merkki vaihtuu siinä positiivisesta negatiiviseksi. (Nimittäjä $\theta\,(1-\theta)$ on positiivinen.) Tämän takia kyseessä on maksimipiste. Derivaatan merkistä nähdään myös, että $\ell(\theta)$ kasvaa välillä $(0, \hat{\theta})$ ja vähenee välillä $(\hat{\theta}, 1)$, joten $\hat{\theta}$ on globaali maksimi. Kuvassa 4.1 esitetään funktion $\ell$ kulkukaavio siinä tilanteessa, kun $n = 7$ toistossa havaitaan $x = 2$ onnistumista.

Tapauksessa $x = n$ 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 $\hat{\theta} = 1 = x/n$. 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 $x = 0$ nädään vastaavasti, että SU-estimaatti on $\hat{\theta} = 0 = x/n$, 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

$$\hat{\theta} = \frac{x}{n} \tag{4.4}$$

eli SU-estimaatti on onnistumisten $x$ 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

$$\sqrt{\frac{1}{n}\,\hat{\theta}\,(1-\hat{\theta})}, \tag{4.5}$$

jossa tuntematon parametrinarvo $\theta$ on korvattu sen estimaatilla $\hat{\theta}$.

Kuvassa 4.2 esitetään binomikokeen uskottavuusfunktio ja logaritminen uskottavuusfunktio kahdella erilaisella otoskoolla. Näissä kuvissa tilanne on valittu siten, että $\hat{\theta} = x/n = 0.2$ 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 (4.5), jolloin otoskoolla $n = 5$ saadaan keskivirhe

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

ja otoskoolla $n = 80$ keskivirhe

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

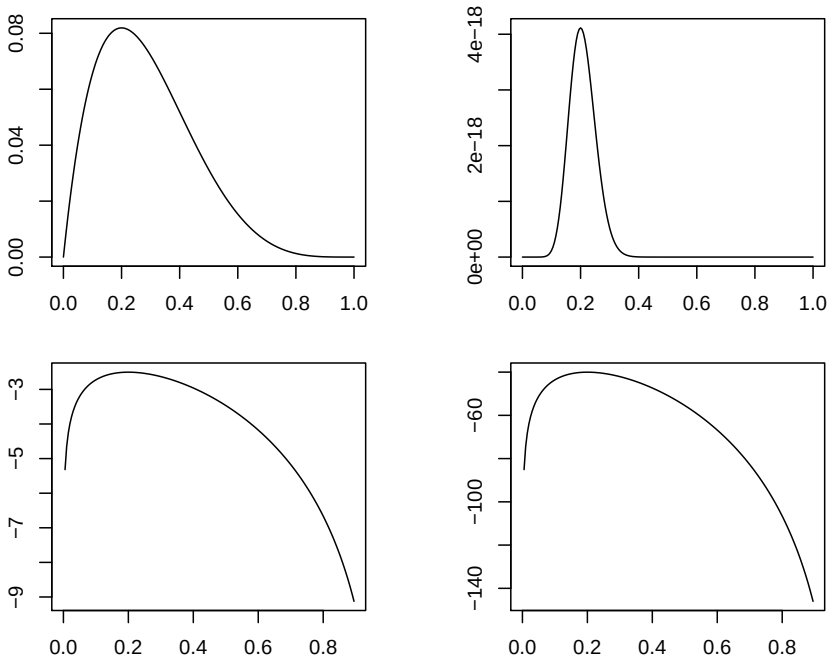
## 4.4 Normaalijakauman parametrien estimointi

Tarkastelemme tilannetta, jossa mallinnamme aineiston $\mathbf{y} = (y_1, \ldots, y_n)$ siten, että vastaavat satunnaismuuttujat $Y_1, \ldots, Y_n$ 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.

Kunkin yksittäisen satunnaismuuttujan $Y_i$ tiheysfunktio on

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

**Kuva 4.2** 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.



Tässä exp tarkoittaa eksponenttifunktiota, eli

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

Parametrien $\mu$ ja $\sigma^2$ merkitys on se, että kullakin $i$

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

Parametri $\mu$ 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

$$\begin{aligned} f(\mathbf{y}; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(-\frac{1}{2\,\sigma^2}(y_i - \mu)^2\right). \\ &= (2\,\pi)^{-n/2}\,(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned} \tag{4.6}$$

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 $b$.

Jätetään uskottavuusfunktiosta $2\pi$:n potenssit pois, jolloin kaavasta (4.6) saadaan havaintoa $\mathbf{y}$ vastaavalle logaritmiselle uskottavuusfunktiolle lauseke

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{4.7}$$

Ylläolevassa kaavassa voidaan neliöiden summa hajottaa kahteen osaan (harjoitustehtävä)

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n\,(\bar{y} - \mu)^2, \tag{4.8}$$

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.9}$$

Tämä huomio helpottaa SU-estimaattien löytämistä.

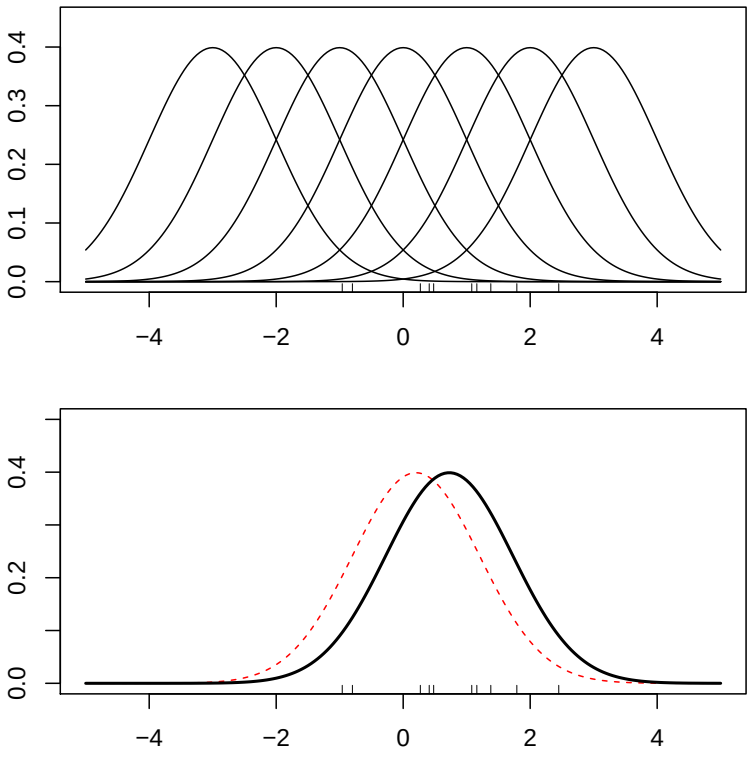
### 4.4.1 Varianssi tunnettu

Jos normaalijakaumaperheessä varianssi $\sigma^2$ on tunnettu luku, niin mallissa on jäljellä vain yksi tuntematon parametri $\mu$. Kuvassa 4.3 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää nyt valita kuvaamaan $x$-akselille lyhyillä viivoilla merkittyä aineistoa.

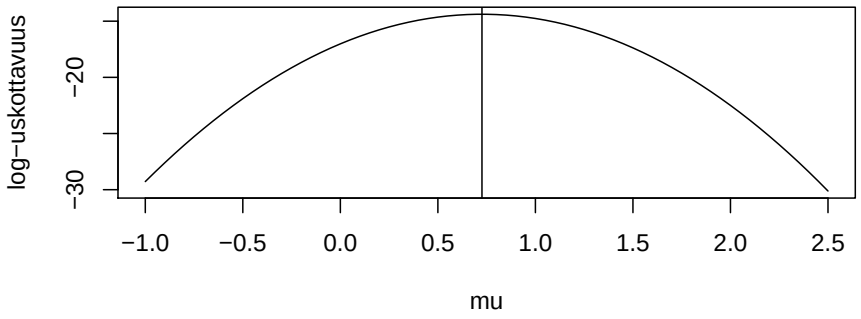
Logaritminen uskottavuusfunktio on kaavojen (4.7) ja (4.8) mukaan

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

**Kuva 4.3** Parametrin $\mu$ estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun $\sigma^2$ on tunnettu luku (tässä $\sigma^2 = 1$). Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, 1)$ tiheysfunktioita muutamilla eri parametrin $\mu$ 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 4.4** Parametrin $\mu$ logaritminen uskottavuusfunktio. SU-estimaatti on merkitty pystyviivalla.




Tässä vakioksi merkitty termi ei riipu $\mu$:sta. 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 4.4 kuvan 4.3 aineistolle.

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

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{4.10}$$

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

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

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

### 4.4.2 Molemmat parametrit tuntemattomia

Nyt molemmat parametrit $\mu$ ja $\sigma^2$ ovat tuntemattomia, joten satunnaisvektorin $\mathbf{Y}$ jakauman kiinnittämiseksi pitäisi tuntea parametrivektorin $\boldsymbol{\theta} = (\mu, \sigma^2)$ arvo. Kuvassa 4.5 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää jälleen valita kuvaamaan $x$-akselille lyhyillä viivoilla merkittyä aineistoa.

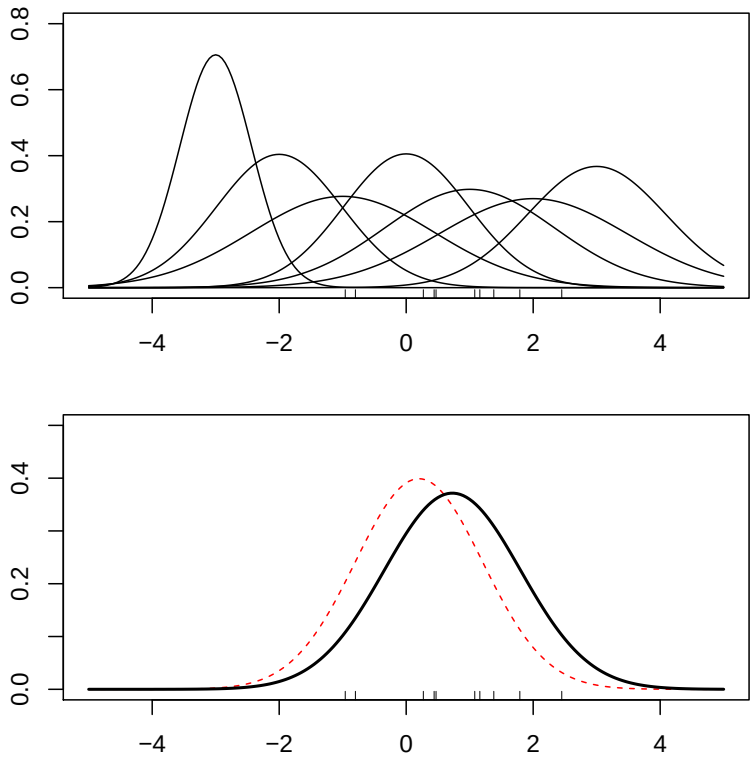
Logaritminen uskottavuusfunktio on kaavojen (4.7) ja (4.8) mukaan

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

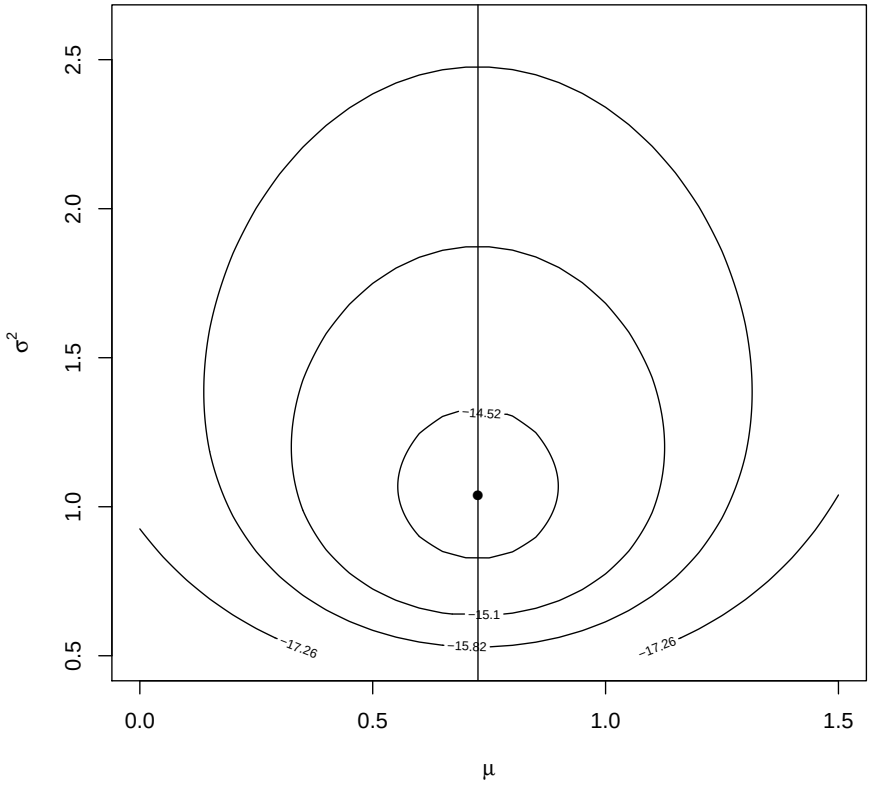
Logaritminen uskottavuusfunktio on esitetty kuvassa 4.6 kuvan 4.3 aineistolle.

Logaritminen uskottavuusfunktio riippuu $\mu$:n arvosta vain sen viimeisen termin kautta. Oli varianssiparametrin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion

**Kuva 4.5** Parametrin $(\mu, \sigma^2)$ estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun sekä $\mu$ että $\sigma^2$ ovat tuntemattomia. Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, \sigma^2)$ tiheysfunktioita muutamilla eri parametrivektorin $(\mu, \sigma^2)$ 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.



**Kuva 4.6** Parametrivektorin $(\mu, \sigma^2)$ 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.



$\mu \mapsto \ell(\mu, \sigma^2)$ maksimoi arvo $\hat{\mu} = \bar{y}$. Tämän ansiosta maksimointi saadaan palautettua yhdestä muuttujasta riippuvan funktion $u$ maksimointitehtäväksi, jossa

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

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 $(\mu, \sigma^2)$ SU-estimaatti on

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{4.11}$$

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 (harjoitustehtävä)

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

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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{4.12}$$

Sitä vastaava estimaattori

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{4.13}$$

on harhaton (varianssiparametrille $\sigma^2$), 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 $N(\mu, \sigma^2)$ noudattavia

satunnaismuuttujia, niin tällöin

$$\bar{Y} \text{ ja } S^2 \text{ ovat riippumattomia,} \tag{4.14}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{1}{n}\,\sigma^2), \tag{4.15}$$

$$\frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}. \tag{4.16}$$

Tässä $\chi^2_{n-1}$ 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 (4.15) 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 $\mu$, ja populaation varianssi $\sigma^2$ on ns. *haittaparametri* (engl. *nuisance parameter*), joka tarvitaan mallin spesifioimiseksi, mutta jonka arvosta ei olla kiinnostuneita. Tässä tapauksessa parametrin $\mu$ estimaatti on otoskeskiarvo $\bar{y}$. Vastaavan estimaattorin $\bar{Y}$ (otantajakauman) varianssi on $\sigma^2/n$. Kun tähän kaavaan sijoitetaan tuntemattoman populaatiovarianssin $\sigma^2$ 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 (4.12) neliöjuuri, eli

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.17}$$

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

$$\frac{1}{\sqrt{n}}\, s = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4.18}$$

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

Jos odotusarvo on tuntematon, niin myös muulloin kuin normaalijakautuneen populaation tapauksessa populaation varianssia usein estimoidaan otosvarianssilla $s^2$ (4.12), jota vastaava estimaattori $S^2$ (4.13) on populaation varianssin harhaton estimaattori aina, kun käsitellään satunnaisotosta populaatiosta, jonka varianssi on $\sigma^2$. Populaation keskihajontaa $\sigma = \sqrt{\sigma^2}$ on myös tapana estimoida otoskeskihajonnalla, vaikka vastaava estimaattori $S = \sqrt{S^2}$ ei ole harhaton.

## 4.5 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 $g(y; \boldsymbol{\theta})$.

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

$$\mu_k(\boldsymbol{\theta}) = EY^k = \begin{cases} \sum_y y^k\, g(y; \boldsymbol{\theta}) & \text{jos jakauma on diskreetti,} \\ \int y^k\, g(y; \boldsymbol{\theta})\, \mathrm{d}y & \text{jos jakauma on jatkuva.} \end{cases} \tag{4.19}$$

Tutuille jakaumille momenttien kaavat tunnetaan. Momenttia $\mu_k(\theta)$ voidaan estimoida $k$:nnella otosmomentilla

$$m_k(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^k, \tag{4.20}$$

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

Momenttimenetelmässä estimaatti (tai estimaattori) muodostetaan ratkaisemalla yhtälöryhmästä

$$\begin{cases} \mu_1(\boldsymbol{\theta}) & = m_1 \\ \mu_2(\boldsymbol{\theta}) & = m_2 \\ & \vdots \\ \mu_r(\boldsymbol{\theta}) & = m_r \end{cases} \tag{4.21}$$

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.

**Esimerkki 4.3** (Eksponenttijakauman parametrin estimointi momenttimenetelmällä) Eksponenttijakaumaa noudattava satunnaismuuttuja $Y$ voi saada kaikkia positiivisia reaaliarvoja, ja sillä on tiheysfunktio

$$g(y;\lambda) = \lambda \, \mathrm{e}^{-\lambda \, y}, \qquad y > 0. \tag{4.22}$$

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

$$EY = \frac{1}{\lambda}.$$

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ eksponenttijakaumasta $\mathrm{Exp}(\lambda)$, ja havaitut arvot ovat $y_1, \ldots, y_n$. Koska parametreja on vain yksi, momenttimenetelmässä tarvitaan vain yksi yhtälö

$$EY = \frac{1}{\lambda} = m_1 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

Momenttimenetelmän mukainen parametrin $\lambda$ estimaatti on

$$\hat{\lambda} = \frac{1}{\bar{y}}.$$

Vaihtoehtoisesti eksponenttijakauma $\mathrm{Exp}(\lambda)$ voitaisiin parametroida sen odotusarvolla $\theta = 1/\lambda$. Momenttimenetelmä antaa tälle parametrille estimaatin

$$\hat{\theta} = \bar{y}.$$

Eksponenttijakauman parametrille voi helposti johtaa myös SU-estimaatin kummalla tahansa parametroinnilla. Tässä esimerkissä momenttimenetelmä antaa samat estimaatit kuin SU-menetelmä, mutta yleisesti ottaen nämä menetelmät voivat tuottaa erilaiset estimaatit. △

# Luku 5

# Luottamusvälit ja luottamusjoukot

## 5.1 Johdanto

On epärealistista ajatella, että piste-estimaatilla löydettäisiin juuri oikea parametrinarvo. Siksi on tarpeen arvioida piste-estimaatin tarkkuutta. Edellisessä luvussa tätä tarkoitusta varten laskettiin keskivirheitä. Tässä luvussa parametriavaruudesta rajataan joukko (mielellään mahdollisimman pieni joukko), joka sisältää todellisen parametrinarvon suurella todennäköisyydellä (toistetussa otannassa). Tällöin puhutaan luottamusjoukosta.

Jos estimoitava parametri on yksiulotteinen ja jos luottamusjoukko on väli, niin silloin sitä kutsutaan luottamusväliksi.

Useissa tilastollisissa malleissa joudutaan tyytymään likimääräisiin luottamusväleihin (tai -joukkoihin).

Luottamusvälien sijasta on joskus mielekästä tarkastella aivan muunlaisia välejä, esim. enuustevälejä.

## 5.2 Luottamusjoukon määritelmä

Tarkastelemme frekventististä tilastollista mallia eli jakaumaperhettä

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

sekä satunnaisvektoria $\mathbf{Y}$, joka noudattaa jakaumaa $f(\mathbf{y};\theta)$ jollakin parametrinarvolla $\theta \in \Theta$.

**Määritelmä 5.1** (Luottamusjoukko)**.** Olkoon $0 < \alpha < 1$ jokin luku. Aineistosta riippuva $\Theta$:n osajoukko $A(\mathbf{y})$ on parametrin $\tau = k(\theta)$ *luottamusjoukko* (engl. *confidence set*) *luottamustasolla* $1-\alpha$ (engl. *confidence level; confidence coefficient*), mikäli vastaava satunnaisvektorista $\mathbf{Y}$ laskettu joukko toteuttaa ehdon

$$P_\theta\left(\tau \in A(\mathbf{Y})\right) \geq 1-\alpha, \qquad \text{kaikilla } \theta \in \Theta. \tag{5.1}$$

Luottamusväli on luottamusjoukko, joka on lukusuoran väli, joten se voidaan määritellä seuraavasti.

**Määritelmä 5.2** (Luottamusväli)**.** Aineistosta laskettua väliä $[L, U]$ sanotaan skalaariparametrin $\tau = k(\theta)$ luottamusväliksi (engl. *confidence interval, CI*) luottamustasolla $1 - \alpha$, jos vastaaville satunnaisille välin päätepisteille $L(\mathbf{Y})$ ja $U(\mathbf{Y})$ pätee

$$P_\theta(L(\mathbf{Y}) \leq \tau \leq U(\mathbf{Y})) \geq 1 - \alpha \tag{5.2}$$

Huomautuksia

- Tässä (kuten tilastollisissa testeissä) $\alpha$ on virhetodennäköisyys. Se on tavallisesti pieni luku, ja tyypillisin valinta on $\alpha = 0.05$, jolloin luottamustaso on $1 - \alpha = 0.95$, eli 95%. Tällöin usein sanotaan lyhyesti, että $A(\mathbf{y})$ on parametrin $\tau$ 95%:n luottamusjoukko. Toinen tavanomainen valinta on $\alpha = 0.01$, mikä vastaa luottamustasoa 99%.

- Satunnaisuus viittaa aineistoa vastaavan satunnaisvektorin $\mathbf{Y}$ jakaumaan (tai toistettuun otantaan).

- Frekventistisessä päättelyssä parametri $\theta$ ei ole satunnainen, vaan kiinteä. Havaintoaineistosta laskettu luottamusjoukko $A(\mathbf{y})$ joko sisältää tai ei sisällä todellista parametrinarvoa $\tau = k(\theta)$, eikä tähän sisälly enää mitään satunnaisuuta. Tämän takia tarvitaan taas uusi termi: luottamusjoukko, luottamusväli. (Ei voida puhua esim. todennäköisyysvälistä.)

- Tahtoisimme luottamusjoukon olevan jollakin tavalla pieni. Koko parametriavaruus $A(\mathbf{y}) = \Theta$ olisi minkä tahansa tason $1 - \alpha$ luottamusjoukko mallin parametrille $\theta$, mutta tämä triviaali luottamusjoukko ei kiinnosta ketään.

- Kaikkein mieluiten konstruoisimme luottamusjoukon sillä tavalla, että kaavassa (5.1) peittotodennäköisyys (engl. *coverage probability*)

  $$P_\theta\left(\tau \in A(\mathbf{Y})\right)$$

  olisi tasan $1-\alpha$ koko parametriavaruudessa. Tietyissä yksinkertaisissa malleissa tämä on mahdollista. Toisinaan tätä vaatimusta on kuitenkin mahdotonta toteuttaa, ja sen takia määritelmässä sallitaan myös epäyhtälö.

## 5.3 Saranasuure

Jos havaintojen jakauma on jatkuva ja jos parametriavaruus on jatkuva, niin eräissä tärkeissä malleissa on mahdollista löytää luottamusjoukko, jolla on tarkalleen haluttu peittotodennäköisyys $1 - \alpha$. Konstruktioon tarvitaan ns. saranasuure.

**Määritelmä 5.3** (Saranasuure)**.** Parametrin $\tau = k(\theta)$ ja satunnaisvektorin $\mathbf{Y}$ funktiota, jonka jakauma ei riipu parametrinarvosta, kutsutaan *saranasuureeksi* (tai napamuuttujaksi) (engl. *pivotal quantity, pivot*) parametrille $\tau$.

**Esimerkki 5.1** Jos $Y_1, \ldots, Y_n$ on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, ja varianssiparametri $\sigma^2$ on tunnettu luku, niin tällöin

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

josta nähdään, että

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

joten $Z$ on saranasuure. Huomaa, että se ei ole tunnusluku, koska sen arvoa ei pystytä laskemaan, jos tunnetaan $\mathbf{Y}$:n arvo, mutta ei parametrinarvoa $\theta = \mu$ (tässä $\sigma^2$ on tunnettu luku). △

Jos normaalijakauman varianssi on tuntematon, niin osoittautuu että analogisesti muodostetulla saranasuureella on ns. $t$-jakauma tietyllä vapausasteparametrilla $\nu$. Nämä $t$-jakaumat ovat sellainen jakaumaperhe, jossa jokaista positiivista reaalilukua $\nu > 0$ kohti on olemassa vastaava jakauma $t_\nu$.

## 5.4 Ala- ja yläkvantiilit

Luottamusvälin konstruoimiseen tarvitsemme saranasuureen jakauman ns. kriittisiä arvoja, jotka lasketan sen kvantiilifunktion avulla. Kvantiilifunktion arvoja kutsutaan myös (jakauman) kvantiileiksi tai fraktiileiksi. Määrittelemme kvantiilifunktion vain jatkuvassa tapauksessa.

Olkoon satunnaismuuttujalla $X$ jatkuva jakauma. Oletamme lisäksi, että sen *kertymäfunktio* (engl. *(cumulative) distribution function*

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(v)\,\mathrm{d}v$$

on aidosti kasvava jollakin välillä $(a, b)$, joka sisältää tämän jakauman koko todennäköisyysmassan, ts. $P(X \in (a, b)) = 1$. Tässä yhteydessä salimme välin $(a, b)$ päätepisteille myös arvot $a = -\infty$ tai $b = \infty$. Esimerkiksi

- standardinormaalijakaumalle $N(0, 1)$ tai $t$-jakaumalle $t_\nu$ tällainen väli on $(-\infty, \infty)$;
- khiin neliön jakaumalle $\chi^2_\nu$ tällainen väli on $(0, \infty)$.

Edeltävillä oletuksilla millä tahansa $0 < u < 1$ on olemassa yksikäsitteinen piste $x \in (a, b)$ siten, että

$$F_X(x) = u \tag{5.3}$$

Tämän yhtälön ratkaisua $x = q(u) \in (a, b)$ kutsutaan satunnaismuuttujan $X$ (tai sen jakauman) *u-kvantiiliksi* $q$ (engl. *u quantile*) eli sen *kvantiilifunktion* (engl. *quantile function*) arvoksi pisteessä $0 < u < 1$. Huomaa, että

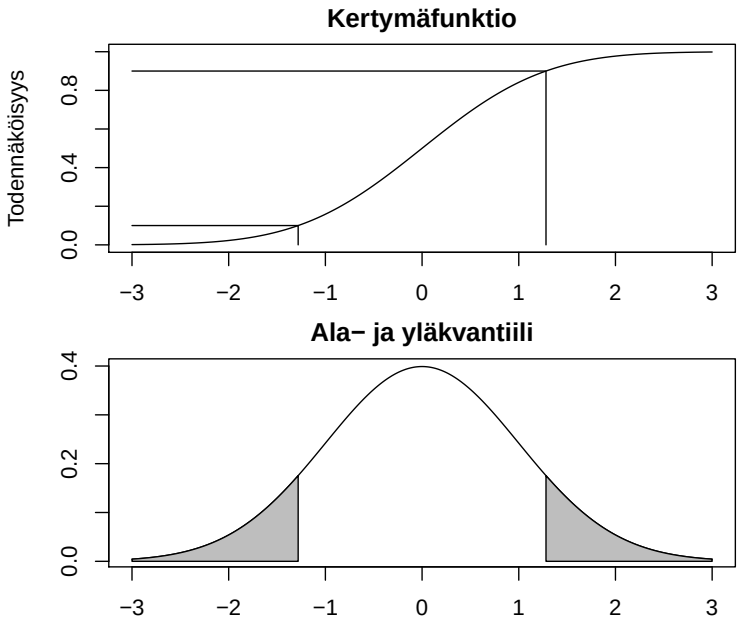
$$q(u) = x \quad \text{eli} \quad F_X(x) = u$$

täsmälleen silloin kuin

$$P(X \leq x) = P(X < x) = u \quad \text{ja} \quad P(X > x) = P(X \geq x) = 1 - u.$$

Ylläolevia todennäköisyyksiä kutsutaan usein *häntätodennäköiyyksiksi* (engl. *tail probability*) tai häntäalueen todennäköisyyksiksi (engl. *tail-area probability*). Jatkuvien jakaumien kohdalla voidaan puhua häntäalueiden pinta-aloista, ks. esim. kuvaa 5.1.

**Kuva 5.1** Standardinormaalijakauman $N(0,1)$ kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$. Kummankin varjostetun häntäalueen pinta-ala on $u$.




**Määritelmä 5.4** (Ala- ja yläkvantiilit)**.** Sellaista pistettä, josta oikealle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman $u$-yläkvantiiliksi (engl. *upper u quantile*).

Sellaista pistettä, josta vasemmalle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman $u$-kvantiiliksi tai $u$-alakvantiiliksi.

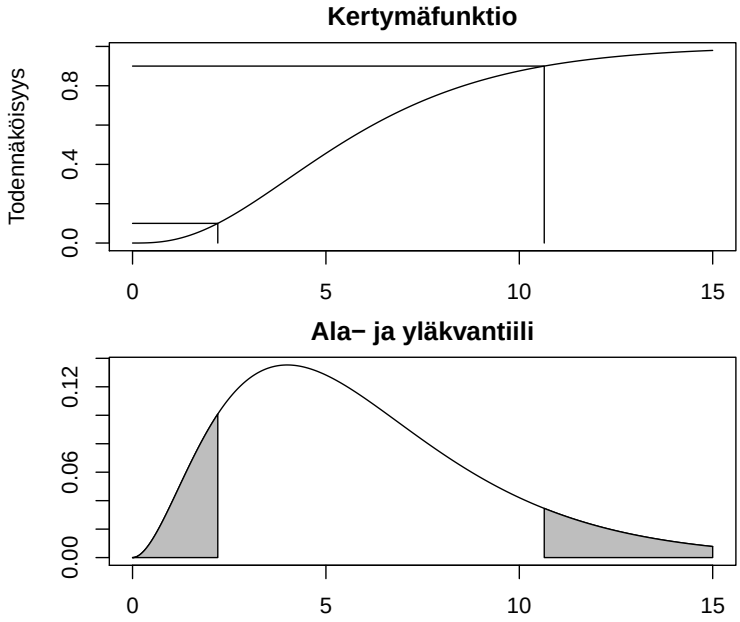
Huomautuksia:

- Termit alakvantiili ja yläkvantiili eivät ole kovin yleisessä käytössä; yleensä käytetään pidempiä ilmaisuja.
- Kvantiilifunktion $q$ avulla lausuttuna $u$-kvantiili eli $u$-alakvantiili on $q(u)$ ja $u$-yläkvantiili on $q(1-u)$.
- Kvantiileja kutsutaan myös fraktiileiksi, ja usein luku $u$ ilmaistaan prosenteissa. Tällöin alakvantiilille käytetään myös nimeä persentiili tai prosenttipiste.

Kuvassa 5.1 havainnollistetaan ala- ja yläkvantiileja sekä vastaavia häntäalueita standardinormaalijakaumalle $N(0,1)$, ja kuvassa 5.2 taas tietylle khiin neliön jakaumalle.

Vanhemmissa tilastotieteen oppikirjoissa on liitteenä laajat taulukot esim. standardinormaalijakauman, t-jakauman ja khiin neliön jakauman kvantiilifunktioista (tai kriittisistä pisteistä). Tällaiset taulukot ovat nykyaikana tarpeettomia. Tilastollisilla ohjelmistoilla saadaan nykyään (tietokoneella tai jopa älypuhelimella) vaivattomasti selville päättelyssä tarvittavat ala- ja yläkvantiilit. Niitä löytyy myös monilta verkkosivuilta, kuten esim.

**Kuva 5.2** Khiin neliön $\chi^2_\nu$ kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$ ja vapausasteluku $\nu = 6$. Kummankin varjostetun häntäalueen pinta-ala on $u$.




`http://www.statsoft.com/textbook/distribution-tables/`

Esimerkiksi R-ohjelmistolla standardinormaalijakauman alakvantiilit pisteissä 0.1, 0.05, 0.025, 0.01 ja 0.005 saadaan laskettua komennoilla

```
u <- c(0.1, 0.05, 0.025, 0.01, 0.005)
qnorm(u)

## [1] -1.282 -1.645 -1.960 -2.326 -2.576
```

ja yläkvantiilit samoissa pisteissä komennolla

```
qnorm(u, lower = FALSE)

## [1] 1.282 1.645 1.960 2.326 2.576
```

Vastaavasti $t$-jakauman ala- ja yläkvantiilit saadaan laskettua (annetulla $\nu$:n arvolla) komennoilla

```
nu <- 6
qt(u, df = nu)
```

```
## [1] -1.440 -1.943 -2.447 -3.143 -3.707

qt(u, df = nu, lower = FALSE)

## [1] 1.440 1.943 2.447 3.143 3.707
```

ja khiin neliön jakauman ala- ja yläkvantiilit komennoilla

```
qchisq(u, df = nu)

## [1] 2.2041 1.6354 1.2373 0.8721 0.6757

qchisq(u, df = nu, lower = FALSE)

## [1] 10.64 12.59 14.45 16.81 18.55
```

Jos jakauma on symmetrinen (ts. sen tiheysfunktio on parillinen funktio), niin tällöin $u$-alakvantiili on $u$-yläkvantiilin vastaluku, sillä symmetriselle jakaumalle

$$q(1-u) = -q(u) \quad \text{kaikille } 0 < u < 1,$$

vrt. kuva 5.1. Tämän takia symmetrisille jakaumille ei tarvita kuin toista jakauman häntää vastaavat kvantiilit. Näille käytetään usein lyhyitä merkintöjä. Tässä monisteessa

$$z_u \quad \text{on } N(0,1)\text{-jakauman } u\text{-yläkvantiili} \tag{5.4}$$

$$t_\nu(u) \quad \text{on } t_\nu\text{-jakauman } u\text{-yläkvantiili.} \tag{5.5}$$

*Varoitus:* Merkinnät ovat eri lähteissä erilaisia. Useissa kirjoissa $z_\alpha$ tarkoittaa $N(0,1)$-jakauman $u$-kvantiilia eikä $u$-yläkvantiilia. Joissakin lähteissä $z_\alpha$ tarkoittaa $N(0,1)$-jakauman $\alpha/2$-yläkvantiilia. Vapausasteluvun merkintä $t$-jakauman yhteydessä on kirjavaa.

## 5.5 Luottamusjoukon muodostaminen saranasuureen avulla

Olkoon nyt $h(\tau, \mathbf{Y})$ saranasuure parametrille $\tau = k(\theta)$. Määritelmän mukaan tämä tarkoittaa sitä, että saranasuureen jakauma on sama riippumatta siitä, mikä on parametrinarvo $\theta \in \Theta$. Oletamme, että tämä jakauma on jatkuva, ja merkitsemme sen kvantiilifunktiota kirjaimella $q$.

Mikäli $0 < \alpha < 1$ on annettu, ja valitsemme luvut $\alpha_1 > 0$ ja $\alpha_2 > 0$ siten, että

$$\alpha = \alpha_1 + \alpha_2$$

niin tällöin

$$P_\theta\left[q(\alpha_1) \le h(\tau, \mathbf{Y}) \le q(1-\alpha_2)\right] = 1 - \alpha, \qquad \text{kaikilla } \theta$$

sillä alempaan jakauman häntään jää saranasuureen jakauman todennäköisyysmassasta osuus $\alpha_1$ ja ylempään häntään osuus $\alpha_2$. Tästä näemme, että

$$A(\mathbf{y}) = \{\tau : q(\alpha_1) \leq h(\tau, \mathbf{y}) \leq q(1-\alpha_2)\} \tag{5.6}$$

on parametrin $\tau$ luottamusjoukko (luottamus-)tasolla $1-\alpha$. Rajankäynnillä ($\alpha_1 \to 0$ tai $\alpha_2 \to 0$) saadaan vielä seuraavat luottamusjoukot

$$A(\mathbf{y}) = \{\tau : h(\tau, \mathbf{y}) \leq q(1-\alpha)\}$$
$$A(\mathbf{y}) = \{\tau : q(\alpha) \leq h(\tau, \mathbf{y})\}$$

Se miten virhetodennäköisyys $\alpha$ jaetaan alemmalle ja ylemmälle saranasuureen jakauman hännälle riippuu siitä, minkälainen joukko parametrille saadaan ratkaisemalla ko. epäyhtälöt: epäyhtälöpari (5.6) tai nämä yksittäiset epäyhtälöt. Yleisin valinta on

$$\alpha_1 = \alpha_2 = \alpha/2,$$

ja tällöin voidaan puhua tasahantäisestä (engl. *equal tail*) luottamusvälistä.

Jotta luottamujoukko ei olisi tarpeettoman suuri, niin saranasuureen pitäisi olla järkevä. Se ei saisi (jossain mielessä) hukata aineistoon sisältyvää informaatiota parametrin todellisesta arvosta. Normaalijakaumamallin tapauksessa tulemme käyttämään tällaisia järkeviä saranasuureita.

## 5.6 Luottamusvälejä normaalijakaumamallissa

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Ts. satunnaismuuttujat $Y_i$ ovat riippumattomia, ja niillä on kaikilla normaalijakauma $N(\mu, \sigma^2)$. Muodostamme saranasuureen avulla luottamusvälin parametrille $\mu$ kahdessa tilanteessa.

1) Kun varianssiparametri on tunnettu, jolloin mallin parametri on $\mu$.

2) Kun sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Lopuksi muodostamme vielä luottamusvälin varianssiparametrille $\sigma^2$.

### 5.6.1 Odotusarvon luottamusväli, kun varianssi on tunnettu

Tämä on se tapaus, jossa luottamusvälin muodostaminen on helpointa ymmärtää. Valitettavasti tätä tapausta ei käytännössä tarvita juuri koskaan, sillä hyvin harvoin normaalijakauman varianssi on tunnettu mutta sen odotusarvo on tuntematon.

Käytämme saranasuuretta (vrt. esimerkki 5.1)

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \tag{5.7}$$

joka noudattaa standardinormaalijakaumaa $N(0,1)$. Jos $0 < \alpha < 1$ on annettu, ja luvut $\alpha_1 > 0$ ja $\alpha_2 > 0$ on valittu niin, että $\alpha_1 + \alpha_2 = \alpha$, niin todennäköisyydellä $1-\alpha$ pätee epäyhtälöpari

$$q(\alpha_1) \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq q(1-\alpha_2), \tag{5.8}$$

missä $q$ on $N(0,1)$-jakauman kvantiilifunktio.

Merkitään väliaikaisesti

$$q_1 = q(\alpha_1), \qquad \text{ja} \quad q_2 = q(1-\alpha_2),$$

ja ratkaistaan kaksoisepäyhtälö (5.8) $\mu$:n suhteen:

$$\begin{aligned} & q_1 \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le q_2 \\ \Leftrightarrow \quad & q_1 \frac{\sigma}{\sqrt{n}} \le \bar{Y}-\mu \le q_2 \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow \quad & -q_2 \frac{\sigma}{\sqrt{n}} \le \mu-\bar{Y} \le -q_1 \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow \quad & \bar{Y}-q_2 \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y}-q_1 \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Ratkaisu on väli, joten tulokseksi saadaan luottamusväli parametrille $\mu$.

Tässä tapauksessa on tavanomaista jakaa virhetodennäköisyys tasan alemman ja ylemmän saranasuureen jakauman hännän kesken, jolloin valitaan

$$\alpha_1 = \alpha_2 = \frac{\alpha}{2}.$$

Tällöin $N(0,1)$-jakauman symmetrisyyden ja sopimuksen (5.4) mukaan

$$q_1 = q(\alpha/2) = -z_{\alpha/2} \quad \text{ja} \quad q_2 = q(1-\alpha/2) = z_{\alpha/2},$$

joten

$$P_\mu\left(\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.9}$$

Olemme johtaneet parametrin $\mu$ luottamustason $1-\alpha$ luottamusvälin, kun normaalijakaumaa noudattavan populaation varianssi $\sigma^2$ on tunnettu luku, nimittäin

$$[\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \tag{5.10}$$

Sitä kutsutaan toisinaan $z$-luottamusväliksi, jotta se erotettaisiin myöhemmin käsiteltävästä ns. $t$-luottamusvälistä. Nimi $z$ tulee viitejakaumana käytettävästä $N(0,1)$-jakaumasta, jota noudattavaa satunnaismuuttujaa usein merkitään kirjaimella $Z$. Luottamusväli (5.10) on symmetrinen piste-estimaatin $\bar{y}$ suhteen, ja se voidaan ilmoittaa myös kaavalla

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Aikaisemmin opitun mukaisesti $\sigma/\sqrt{n}$ on $\mu$:n piste-estimaatin (eli otoskeskiarvon $\bar{y}$, joka on SU-estimaatti) keskivirhe. Huomaa, että luottamusvälin leveys riippuu luotamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen puolittaa tämän luottamusvälin leveyden.

Luottamusväli (5.10) on kaksisuuntainen (eli kaksitahoinen) (engl. *two-sided*). On myös mahdollista johtaa yksisuuntaiset (engl. *one-sided*) luottamusvälit. Todennäköisyydellä $1-\alpha$ pätee epäyhtälö

$$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le q(1-\alpha) = z_\alpha,$$

ja kun tämä ratkaistaan $\mu$:n suhteen, nähdään että

$$P_\mu\left(\mu \geq \bar{Y} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.11}$$

Toisaalta todennäköisyydellä $1 - \alpha$ pätee myöskin epäyhtälö

$$-z_\alpha = q(\alpha) \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}},$$

ja kun tämä ratkaistaan, nähdään että

$$P_\mu\left(\mu \leq \bar{Y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.12}$$

Ts. seuraavat aineistosta $\mathbf{y}$ lasketut yksisuuntaiset välit ovat luottamustason $1 - \alpha$ luottamusvälejä

$$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \tag{5.13}$$

$$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}] \tag{5.14}$$

### 5.6.2 Aineistosta lasketun luottamusvälin tulkinta

Lasketaan nyt 95% luottamusväli (5.10) (eli kaksisuuntainen $z$-luottamusväli) populaation odotusarvolle $\mu$ käyttämällä kuvan 4.3 aineistoa olettaen, että tiedämme että $\sigma^2 = 1$. (Simuloinnissa käytettiin tätä varianssia.) Käyttämällä tietoja

$$\bar{y} = 0.726, \quad n = 10, \quad z_{0.025} = 1.96$$

saadaan laskettua parametrille $\mu$

- piste-estimaatti 0.73 (eli SU-estimaatti $\bar{y}$)
- estimaatin keskivirhe 0.32 (eli $\sigma/\sqrt{n}$)
- 95%:n luottamusväli $[0.10, 1.35]$ (eli $\bar{y} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$).

Simuloinnissa käytetty todellinen parametrinarvo $\mu = 0.2012$ kuuluu laskettuun luottamusväliin.

R:n peruspaketeissa ei ole toteutettuna $z$-luottamusväliä. Ohjelmiston kehittäjät ovat luultavasti arvioineet, ettei sitä todellisuudessa koskaan tarvita. Tarvittavat laskut saadaan tehtyä esim. seuraavasti.

```
y <- c(1.38, -0.96, 1.08, 0.41, 0.48, 2.45, -0.8, 0.27, 1.79,
    1.16)
n <- length(y)
z <- qnorm(0.05/2, lower = FALSE)
sigma <- 1
sigma/sqrt(n)

## [1] 0.3162
```

```
mean(y) - z * sigma/sqrt(n)

## [1] 0.1062

mean(y) + z * sigma/sqrt(n)

## [1] 1.346
```

Ennen aineiston keräämistä (ts. simulointia) tiedämme, että aineistosta laskettava 95%:n luottamusväli tulee sisältämään todellisen populaation keskiarvon todennäköisyydellä 95%. Sitten aineisto kerättiin (tässä: simuloitiin), ja luottamusväliksi saatiin $[0.10, 1.35]$.

*Kysymys: Voimmeko sanoa, että* $\mu \in [0.10, 1.35]$ *todennäköisyydellä 0.95?*

Pysähdy pohtimaan tätä kysymystä, ja muodosta asiasta oma mielipiteesi ennen kuin luet alla olevan vastauksen!

*Vastaus:*

- Aineistosta laskettu luottamusväli joko sisältää todellisen parametrinarvon tai tai ei sisällä sitä. Emme voi pelkästään aineistoa tarkastelemalla sanoa mitään sen enempää, vaan tätä varten pitäisi tuntea todellinen parametrinarvo.

- Frekventistisessä tilastotieteessä parametri on tuntematon, mutta kiinteä (siis ei-satunnainen). Tämän lähestymistavan puitteissa väite $\mu \in [0.10, 1.35]$ on joko tosi tai epätosi (nyt se on tosi). Tällaisen väitteen todennäköisyys ei taatusti ole 0.95.

Tämä tulkinnallinen vaikeus ei liity kaavaan (5.10), vaan luottamusvälin käsitteeseen. Luottamusvälin määritelmässä todennäköisyys viittaa siihen, että aineistoa pidetään satunnaisvektorina, jolla on jakauma $f(\mathbf{y}; \theta)$. Tällöin luottamusvälin päätepisteet eli tunnusluvut $L(\mathbf{Y})$ ja ovat $U(\mathbf{Y})$ ovat satunnaismuuttujia, ja todennäköisyydellä $1-\alpha$ todellinen parametrinarvo sisältyy satunnaiselle välille $[L(\mathbf{Y}), U(\mathbf{Y})]$.

Tätä tulkintaa voidaan havainnollistaa ajattelemalla toistettua aineistonkeruuta, jota on havainnollistettu kuvassa 5.3. Jos laskemme luottamusvälin (5.10) suurelle määrälle $r$ normaalijakaumasta $N(\mu, \sigma^2)$ simuloituja kokoa $n$ olevia otoksia (jossa $\sigma^2$ on tunnettu)

$$\mathbf{y}_1, \ldots, \mathbf{y}_r,$$

niin saamme $r$ kappaletta luottamusvälejä

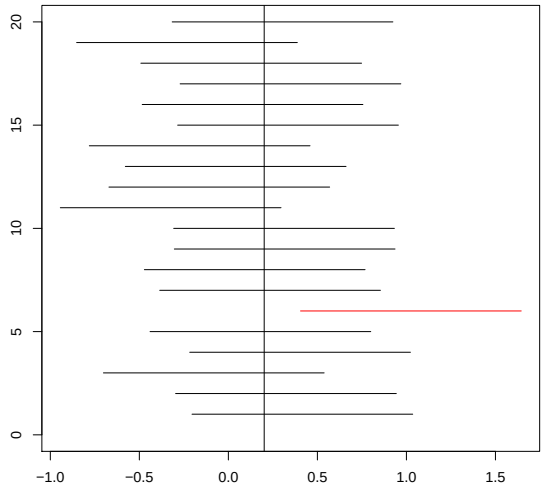
$$[L(\mathbf{y}_1), U(\mathbf{y}_1)], \ldots, [L(\mathbf{y}_r), U(\mathbf{y}_r)].$$

Näistä osapuilleen $r(1-\alpha)$ kappaletta sisältää todellisen parametrinarvon ja $r\alpha$ kappaletta ei sisällä sitä.

*Kysymys:* Hyvä on, *en saa sanoa*, että $\mu \in [0.10, 1.35]$ todennäköisyydellä 0.95. Miten sitten *saan* tulkita aineistosta lasketun luottamusvälin?

*Vastaus:* Aineiston perusteella paras arvauksemme parametrinarvolle on pisteestimaatti 0.73. 95%:n luottamusvälillä $[0.10, 1.35]$ olevat arvot ovat kaikki kohtuullisessa sopusoinnussa havaintojen kanssa. Sekä luottamusvälin leveys että

**Kuva 5.3** 20 kappaletta kaavalla (5.10) laskettua $z$-luottamusväliä jakaumasta $N(\mu, 1)$ simuloiduille, kokoa $n = 10$ oleville aineistoille. Todellinen parametrinarvo on merkitty pystyviivalla. Kun normaalijakauman varianssi tunnetaan, niin luottamusvälin leveys pysyy vakiona.




estimaatin keskivirhe kuvastavat tietomme epävarmuutta parametrinarvosta tämän aineiston valossa. Väli on laskettu sellaisella menetelmällä, joka toistetussa aineistonkeruussa mallin oletukset toteuttavasta populaatiosta sisältäisi todellisen parametrinarvon noin 95% toistoista. Ennen aineistonkeruuta todennäköisyys oli 95%, että siitä laskettava 95%:n luottamusväli tulee sisältämään oikean parametrinarvon (olettaen tietenkin, että populaatio toteuttaa mallioletukset).

### 5.6.3 Odotusarvon luottamusväli, kun varianssi on tuntematon

Tässä tilanteessa sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$. Haluamme muodostaa luottamusvälin odotusarvoparametrille

$$\mu = k(\mu, \sigma^2).$$

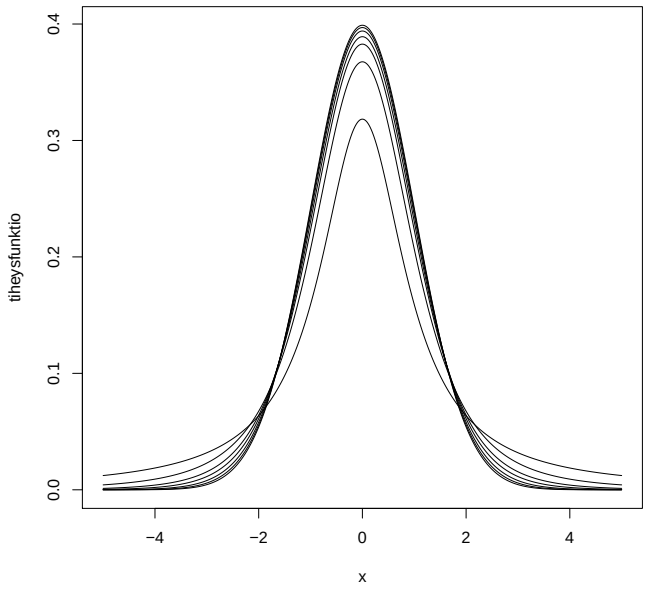
(Tässä funktio $k$ vain palauttaa ensimmäisen argumenttinsa arvon.)

Kun varianssi oli tunnettu, käytimme saranasuuretta

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}.$$

Kun varianssi on tuntematon, $Z$ ei ole saranasuure, koska se riippuu paitsi aineistosta ja kiinnostusparametrista $\mu$, myös haittaparametrista $\sigma^2$. Ajatuksena on kuitenkin matkia mahdollisimman tarkoin aikaisempaa konstruktiota. Koska

**Kuva 5.4** $t_\nu$-jakauman tiheysfunktioita vapausasteluvun $\nu$ arvoilla 1, 3, 6, 10, 20 ja 50. Vertailun vuoksi kuvassa on myös standardinormaalijakauman $N(0,1)$ tiheysfunktio, jota voidaan pitää $t$ jakaumana vapausasteluvulla $\infty$. Tiheysfunktion arvo pisteessä $x = 0$ on sitä suurempi mitä suurempi on vapausasteluku $\nu$. Häntäalueilla järjestys on päinvastainen.




populaation keskihajonta $\sigma$ on tuntematon, sen tilalle sijoitetaan otoskeskihajontaa (4.17) vastaava satunnaismuuttuja $S$. Tässä mallissa satunnaismuuttuja

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{5.15}$$

osoittautuu saranasuureeksi. Sen jakauma on tietty $t$-jakauma.

**Määritelmä 5.5.** Jos $\nu > 0$ ja $Z \sim N(0,1)$ ja $X \sim \chi^2_\nu$ ja $Z$ ja $X$ ovat riippumattomia, niin satunnaismuuttujalla

$$Y = \frac{Z}{\sqrt{X/\nu}}$$

on $t_\nu$-jakauma eli $t$-jakauma vapausasteluvulla $\nu$ (engl. *$t$ distribution with $\nu$ degrees of freedom*).

Määritelmän avulla on mahdollista johtaa $t_\nu$-jakauman tiheysfunktio, mutta tätä kaavaa ei tässä yhteydessä tarvita. Tiheysfunktio osoittautuu parilliseksi funktioksi, joten $t_\nu$-jakauma on symmetrinen. Kuvassa 5.4 esitetään $t_\nu$-jakauman tiheysfunktio muutamilla vapausasteluvun arvoilla. Kun $\nu$ kasvaa, jakauman tiheysfunktio lähestyy standardinormaalijakauman $N(0,1)$ tiheysfunktiota. $t$-jakaumaa kutsutaan myös Studentin $t$-jakaumaksi W. S. Gossetin v.

1908 julkaiseman artikkelin kunniaksi. Tilastotieteilijä W. S. Gosset (1876–1937) työskenteli tuohon aikaan Guinessin panimolla. Panimo oli kieltänyt liikesalaisuuksien suojelemiseksi työntekijöitään julkaisemasta mitään kirjoituksia omalla nimellään, minkä takia Gosset käytti julkaisussa salanimeä Student.

Seuraavaksi tarvitsemme jaksossa 4.4.2 kerrottua tietoa satunnaismuuttujaparin $(\bar{Y}, S^2)$ yhteisjakaumasta (kaavat (4.14), (4.15) ja (4.16)):

- $\bar{Y}$ ja $S^2$ ovat riippumattomia
- $\bar{Y} \sim N(\mu, \frac{1}{n}\,\sigma^2)$,
- $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$.

Edellä mainittujen jakaumatulosten ja $t$-jakauman määritelmän perusteella satunnaismuuttujalla

$$\frac{(\bar{Y}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}} = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$$

on $t$-jakauma vapausasteluvulla $n-1$, mutta sieventämällä nähtiin, että tämä satunnaismuuttuja on sama kuin kaavan (5.15) satunnaismuuttuja $T$.

Olkoon $q$ nyt $t_{n-1}$-jakauman kvantiilifunktio, ja olkoon $0 < \alpha < 1$. Todennäköisyydellä $1-\alpha$ pätee epäyhtälöt

$$q(\alpha_1) \le \frac{\bar{Y}-\mu}{S/\sqrt{n}} \le q(1-\alpha_2),$$

jossa $\alpha_1 > 0$ ja $\alpha_2 > 0$ ovat sellaisia lukuja, joiden summa on $\alpha$. Tästä saadaan ratkaistua väli odotusarvolle $\mu$ aivan samoilla vaiheilla kuin aikaisemmin, ja tulos on

$$\bar{Y} - q(1-\alpha_2)\,\frac{S}{\sqrt{n}} \le \mu \le \bar{Y} - q(\alpha_1)\,\frac{S}{\sqrt{n}}$$

Jos tässä valitaan $\alpha_1 = \alpha_2 = \alpha/2$, ja huomataan, että

$$q(\alpha/2) = -t_{n-1}(\alpha/2) \quad \text{ja} \quad q(1-\alpha/2) = t_{n-1}(\alpha/2),$$

niin päädytään siihen, että

$$P_{(\mu,\sigma^2)}\left(\bar{Y} - t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}} \le \mu \le \bar{Y} + t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}}\right) = 1-\alpha, \qquad (5.16)$$

kaikilla $\mu \in \mathbb{R}$ ja kaikilla $\sigma^2 > 0$.

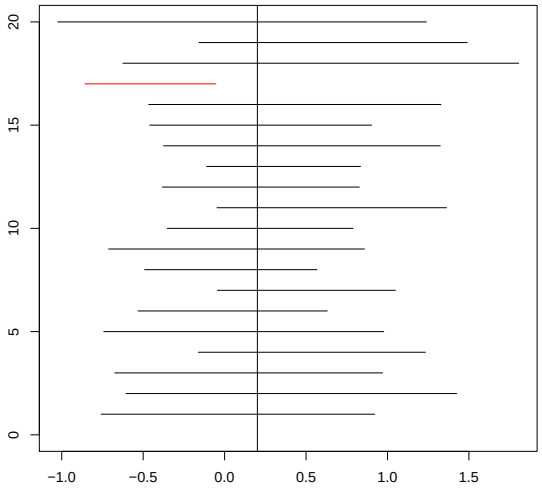
Vastaava aineistosta $\mathbf{y}$ laskettu väli

$$[\bar{y} - t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}] = \bar{y} \pm t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \qquad (5.17)$$

jossa $\bar{y}$ on otoskeskiarvo ja $s$ on otoskeskihajonta, on normaalijakauman odotusarvon $\mu$ luottamusväli luottamustasolla $1-\alpha$. Sitä kutsutaan usein $t$-luottamusväliksi (viitejakauman $t_{n-1}$ mukaan). Huomaa, että $\bar{y}$ on myös parametrin $\mu$ SU-estimaatti ja että $s/\sqrt{n}$ on tämän estimaatin keskivirhe.

Suure $t_{n-1}(\alpha/2)$ lähestyy lukua $z_{\alpha/2}$, kun otoskoko kasvaa. Esimerkiksi luottamustasoa 95% vastaa $\alpha = 0.05$, ja $z_{0.025} = 1.96$. Otoskokoja $n = 50, 100, 200, 500$ ja 1000 vastaavat seuraavat $t$-jakaumaperheen $\alpha/2$-yläkvantiilit

**Kuva 5.5** 20 kappaletta kaavalla (5.17) laskettua $t$-luottamusväliä jakaumasta $N(\mu, \sigma^2)$ simuloiduille, kokoa $n = 10$ oleville aineistoille. Todellinen odotusarvoparametrin arvo on merkitty pystyviivalla. Kun normaalijakauman varianssi on tuntematon, niin luottamusvälin leveys vaihtelee otoksesta toiseen.




```
n <- c(50, 100, 200, 500, 1000)
qt(0.05/2, df = n - 1, lower = FALSE)

## [1] 2.010 1.984 1.972 1.965 1.962
```

Väli (5.17) on symmetrinen piste-estimaatin $\bar{y}$ suhteen. Toisin kuin $z$-luottamusvälin yhteydessä, $t$-luottamusvälin leveys vaihtelee aineistosta toiseen, koska välin leveys määräytyy aineiston otoskeskihajonnasta. Tämän $t$-luottamusvälin leveys riippuu luottamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen karkeasti ottaen puolittaa kaksisuuntaisen $t$-luottamusvälin leveyden (mutta tämä ei pidä paikkaansa tarkalleen).

Kuvassa 5.5 näytetään 20 kappaletta $t$-luottamusvälejä, jotka on laskettu aineistoista, jotka on generoitu tietystä normaalijakaumasta.

Kuten jaksossa 5.6.2 selitettiin, aineistosta lasketulla luottamusvälillä ei ole todennäköisyystulkintaa, vaan se joko sisältää todellisen parametrinarvon tai ei sisällä sitä, emmekä (todellisessa tilanteessa) tiedä kumpi tilanne on kyseessä. Todennäköisyystulkinta vaatii sitä, että tulkitsemme välin päätepisteet satunnaismuuttujiksi tai ajattelemme toistettua aineiston keruuta tai ajattelemme tilannetta, joka vallitsi ennen kuin aineisto kerättiin. Kaikkien luottamusvälin sisällä olevien arvojen voidaan ajatella olevan kohtuullisen hyvin sopusoinnussa aineiston kanssa. Paras arvauksemme on parametrin piste-estimaatti.

**Esimerkki 5.2** Kuvan 4.3 aineistolle

$$\bar{y} = 0.726, \quad s = 1.074, \quad n = 10, \quad t_9(0.025) = 2.262.$$

Parametrin $\mu$ piste-estimaatti on 0.73, sen keskivirhe on 0.34 (kaavalla $s/\sqrt{n}$), ja 95%:n luottamusväli on $[-0.04, 1.50]$.

Tavallisesti luottamusväli lasketaan tietokoneella. R:llä tämä onnistuu seuraavasti

```
y <- c(1.38, -0.96, 1.08, 0.41, 0.48, 2.45, -0.8, 0.27, 1.79,
    1.16)
t.test(y)

##
##  One Sample t-test
##
## data:  y
## t = 2.137, df = 9, p-value = 0.0613
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.04244  1.49444
## sample estimates:
## mean of x
##     0.726
```

Valitettavasti tässä näkyy luottamusvälin selvittämisen kannalta tarpeetonta tietoa; pelkän välin saisi selville antamalla komennon `t.test(y)$conf.int`.

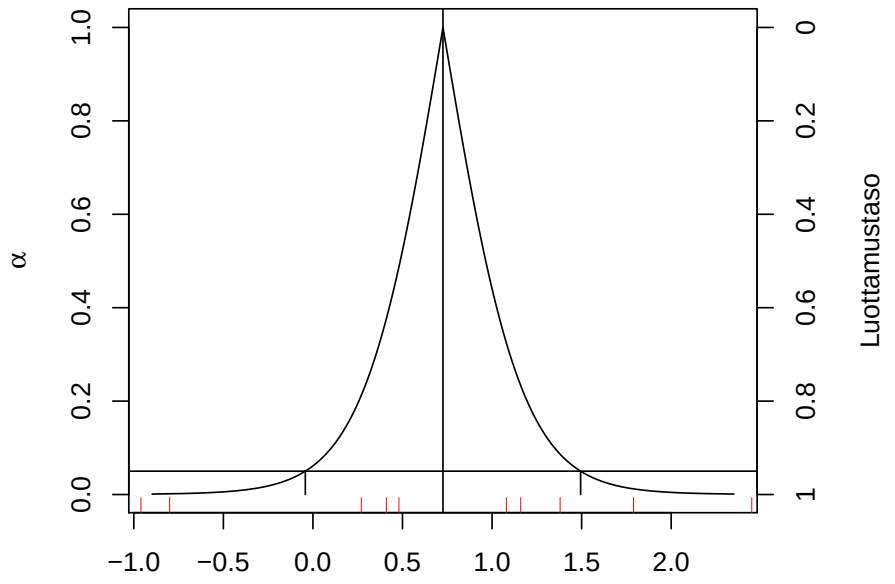
```
t.test(y)$conf.int

## [1] -0.04244  1.49444
## attr(,"conf.level")
## [1] 0.95
```

Jos tahdotaan käyttää muita luottamustasoja kuin 95%, kuten esim. luottamustasoa 99%, niin haluttu luottamustaso pitää antaa `t.test`-funktiolle tyyliin `t.test(y, conf.level = 0.99)`. Valitettavasti `t.test` ei raportoi piste-estimaatin keskivirhettä, mutta sen saa laskettua helposti erikseen seuraavasti.

```
sd(y)/sqrt(length(y))

## [1] 0.3397
```

Mikään ei pakota meitä laskemaan luottamusväliä vain yhdellä luottamustasolla 0.95. Kuvassa 5.6 näytetään luottamusvälin päätepisteet luottamustason funktiona. △

**Kuva 5.6** Kuvan 4.3 aineistolle lasketut parametrin $\mu$ kaksisuuntaiset $t$-luottamusvälit. Piste-estimaatti sekä 95%:n luottamusväli on korostettu pystyviivoilla. Aineisto on esitetty $x$-akselin yläpuolella olevilla pienillä viivoilla.




### 5.6.4 Varianssiparametrin luottamusväli

Oletamme, että sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$. Haluamme muodostaa luottamusvälin varianssiparametrille

$$\sigma^2 = k(\mu, \sigma^2).$$

(Nyt funktio $k$ palauttaa toisen argumenttinsa arvon.)

Käytämme saranasuureena sopivasti skaalattua otosvarianssia, sillä tiedämme, että

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

Jos $q$ on $\chi^2_{n-1}$-jakauman kvantiilifunktio, ja $0 < \alpha < 1$ sekä $\alpha_1 > 0$ ja $\alpha_2 > 0$ ovat lukuja siten, että $\alpha = \alpha_1 + \alpha_2$, niin todennäköisyydellä $1 - \alpha$ pätee

$$q(\alpha_1) \leq \frac{n-1}{\sigma^2} S^2 \leq q(1-\alpha_2)$$

Kun tämä epäyhtälö ratkaistaan muuttujan $\sigma^2$ suhteen, saadaan väli

$$\frac{n-1}{q(1-\alpha_2)} S^2 \leq \sigma^2 \leq \frac{n-1}{q(\alpha_1)} S^2$$

Tässäkin on tapana valita $\alpha_1 = \alpha_2 = \alpha/2$, jolloin varianssiparametrille $\sigma^2$ saadaan kaksisuuntainen tason $1 - \alpha$ luottamusväli

$$[\frac{n-1}{q(1-\alpha/2)} s^2, \frac{n-1}{q(\alpha/2)} s^2], \tag{5.18}$$

jossa $s^2$ on otosvarianssi (joka on varianssiparametrin piste-estimaatti) ja $q$ on $\chi^2_{n-1}$-jakauman kvantiilifunktio. Tämä väli ei ole symmetrinen piste-estimaatin suhteen.

Kuvan 4.3 aineistolle

$$s^2 = 1.1539, \quad n = 10, \quad q(0.025) = 2.7004, \quad q(0.975) = 19.0228,$$

ja näistä luvuista laskettu varianssiparametrin piste-estimaatti on 1.15 ja 95%:n luottamusväli on $[0.55, 3.85]$. Tämä väli sisältää todellisen simuloinnissa käytetyn varianssin $\sigma^2 = 1$.

## 5.7 Likimääräinen luottamusväli

Jos otoskoko $n$ on suuri ja jos piste-estimaattorin $\hat{\tau}(\mathbf{Y})$ otantajakauma on osapuilleen $\tau$-keskinen normaalijakauma, niin tällöin osapuilleen todennäköisyydellä $1 - \alpha$ pätee epäyhtälö

$$-z_{\alpha/2} \leq \frac{\hat{\tau}(\mathbf{Y}) - \tau}{\sqrt{\operatorname{var}_\theta \hat{\tau}(\mathbf{Y})}} \leq z_{\alpha/2}.$$

Kun tämä epäyhtälöpari ratkaistaan parametrin $\tau$ suhteen, saadaan väli

$$\hat{\tau}(\mathbf{Y}) - z_{\alpha/2}\sqrt{\operatorname{var}_\theta \hat{\tau}(\mathbf{Y})} \leq \tau \leq \hat{\tau}(\mathbf{Y}) + z_{\alpha/2}\sqrt{\operatorname{var}_\theta \hat{\tau}(\mathbf{Y})}$$

Tässä estimaattorin otantajakauman keskihajonta $\sqrt{\operatorname{var}_\theta \hat{\tau}(\mathbf{Y})}$ on tavallisesti tuntematon. Jos se korvataan estimaatilla, eli estimaatin $\hat{\tau}$ keskivirheellä, niin päädytään nimellistä (engl. *nominal*) $1 - \alpha$ luottamustasoa vastaavaan (kaksisuuntaiseen) likimääräiseen luottamusväliin

$$\hat{\tau} \pm z_{\alpha/2} \times \text{se}, \tag{5.19}$$

jossa suure (se) on (jollakin järkevällä tavalla laskettu) estimaatin $\hat{\tau}$ keskivirhe.

Koska $z_{0.025} = 1.96$, niin suurella otoskoolla erityisesti

$$\hat{\tau} \pm 2 \times \text{se},$$

on likimääräinen 95%:n luottamusväli. Koska $z_{0.16} = 0.994$, niin suurella otoskoolla

$$\hat{\tau} \pm \text{se},$$

on likimääräinen 68%:n luottamusväli.

Esimerkiksi binomikokeessa onnistumistodennäköisyyden luottamusväli lasketaan tyypillisesti tällä periaatteella. SU-estimaattori

$$\hat{p}(\mathbf{Y}) = \bar{Y}$$

(eli onnistumisten suhteellinen osuus) on harhaton, ja sen varianssi on

$$\operatorname{var}_p \bar{Y} = \frac{1}{n} p(1 - p).$$

Koska estimaattori on keskiarvo $n$ riippumattomasta ja samoin jakautuneesta satunnaismuuttujasta, sen jakaumaa voidaan suurella otoskoolla approksimoida

normaalijakaumalla (todennäköisyyslaskennan keskeisen raja-arvolauseen perusteella). Kun keskivirheelle käytetään kaavaa

$$\mathrm{se} = \sqrt{\frac{1}{n}\,\hat{p}\,(1-\hat{p})},$$

saadaan binomikokeen onnistumistodennäköisyydelle $p$ likimääräinen $1-\alpha$ luottamusväli

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\,\hat{p}\,(1-\hat{p})}, \tag{5.20}$$

joka kerrotaan kaikissa tilastotieteen alkeisoppikirjoissa. Jos $0 < p < 1$ on kiinteä, ja otoskoko $n$ kasvaa rajatta, niin todennäköisyyslaskennan keinoilla voidaan osoittaa, että vastaavan satunnaisen luottamusvälin peittotodennäköisyys lähestyy arvoa $1-\alpha$, joten suurella otoskoolla tämän välin peittotodennäköisyys on suunnilleen $1-\alpha$.

Edellinen asymptoottinen perustelu jättää avoimeksi sen, milloin otoskoko on riittävän suuri. Tämän takia tarkastelemme lähemmin likimääräisen perinteisen likimääräisen luottamusvälin (5.20) ominaisuuksia. Se voi äärellisellä otoskoolla käyttäytyä kummallisella tavalla:

- Sen päätepisteet voivat olla parametriavaruuden ulkopuolella; käytännössä luottamusväliksi pitäisi otta välin (5.20) sekä parametriavaruuden leikkaus.

- Väli surkastuu yhdeksi pisteeksi, jos koesarjassa ei joko onnistuta yhtään kertaa tai jos ei epäonnistuta yhtään kertaa; parametriavaruuden reunojen lähellä tätä väliä ei kannata käyttää.

Kuvassa 5.7 otoskoko on $n = 20$. Siinä esitetään eri onnistumisten lukumäärää $0 \le k \le n$ vastaavat $n+1$ mahdollista luottamusväliä laskettuna kaavalla (5.20). Kuvassa on myös piirretty luottamusvälin todellinen peittotodennäköisyys

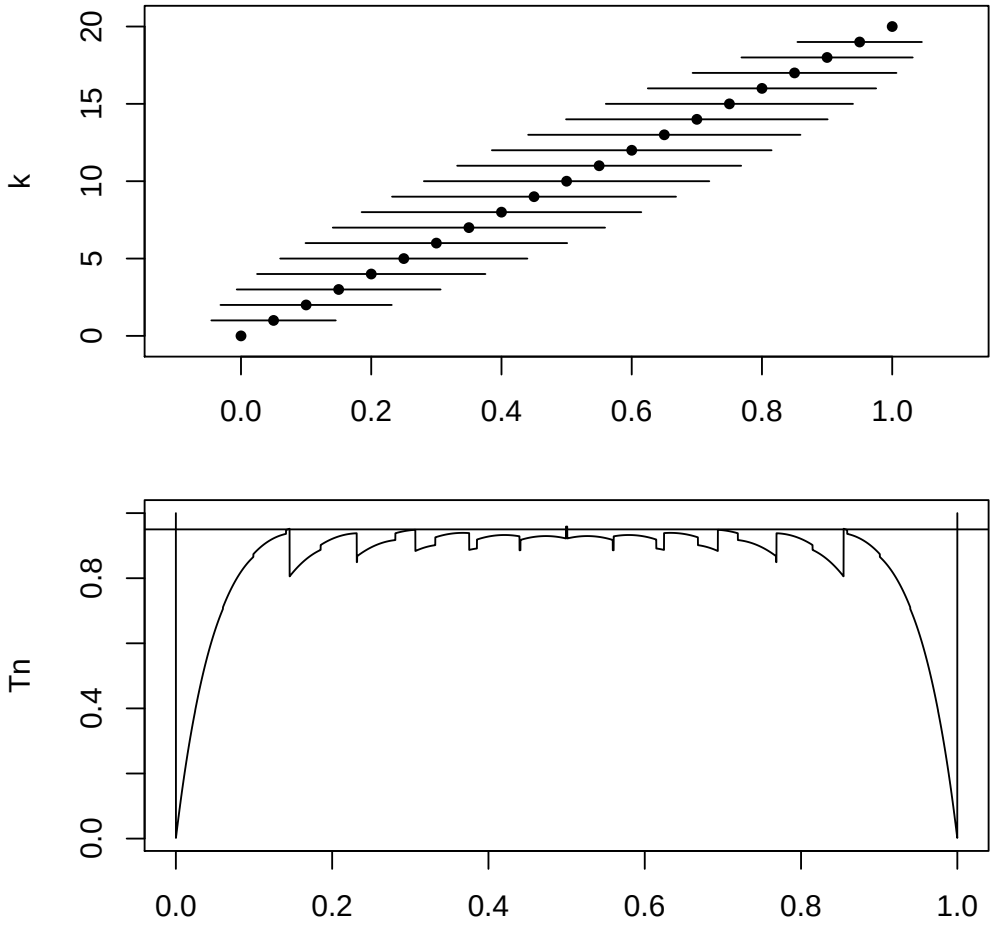
$$P_p(L(\mathbf{Y}) \le p \le U(\mathbf{Y})).$$

Kuvasta näemme, että tällä pienehköllä otoskoolla tämän luottamusvälin todellinen peittotodennäköisyys on melkein koko parametriavaruudessa paljon pienempi kuin nimellinen peittotodennäköisyys. Ainakaan otoskoolla $n = 20$ tätä perinteistä likimääräistä luottamusväliä ei pitäisi käyttää.

## 5.8 Muita luottamusvälejä binomikokeessa

Likimääräisen luottamusvälin (5.20) todellinen peittotodennäköisyys (kun väli tulkitaan satunnaiseksi) käyttäytyy millä tahansa äärellisellä otoskoolla $n$ huonosti joissakin parametriavaruuden pisteissä. Parametriavaruuden reunojen lähellä tämän välin peittotodennäköisyys romahtaa nollaan, koska itse väli surkastuu kummallakin rajalla pisteeksi. Tämän lisäksi todellinen peittotodennäköisyys voi olla selvästi nimellistä tasoa pienempi muuallakin vielä suurehkolla otoskoolla, ks. artikkelia Brown, Cai ja DasGupta [1]. Nämä kirjoittajat toteavat tästä luottamusvälistä seuraavaa:

**Kuva 5.7** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat luottamusvälit (5.20), kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys todellisen onnistumistodennäköisyyden $p$ funktiona. Nimellinen luottamustaso on osoitettu vaakaviivalla.



> ... the performance of this standard interval is persistently chaotic and unacceptably poor. Indeed its coverage properties defy conventional wisdom. The performance is so erratic and the qualifications given in the influential texts are so defective that the standard interval should not be used.

Newcombe [2] vertaa empiirisesti seitsämää erilaista mentelmää luottamusvälin laskemiseksi, ja hän käyttää vertailussa peittotodennäköisyyden lisäksi muitakin kriteereitä. Newcombe kommentoi tätä traditionaalista luottamusväliä (ja sen parannusta, jossa käytetään jatkuvuuskorjausta) seuraavasti,

> ... it is strongly recommended that intervals calculated by these methods should no longer be acceptable for the scientific literature

Nämä neuvot on syytä ottaa huomioon. Älkää käyttäkö perinteistä likimääräistä luottamusväliä (5.20) omissa töissänne.

Mainituissa artikkeleissa käydään läpi monta vaihtoehtoehtoista tapaa muodostaa luottamusväli onnistumistodennäköisyydelle. Esimerkiksi Wilsonin v. 1927 ehdottama luottamusväli osoittautuu edellistä selvästi paremmaksi. Myös se perustuu siihen approksimaatioon, että suureella

$$\frac{\hat{p}(\mathbf{Y}) - p}{\sqrt{\mathrm{var}_p(\hat{p}(\mathbf{Y}))}} = \frac{\hat{p}(\mathbf{Y}) - p}{\sqrt{\frac{1}{n}\,p(1-p)}}$$

on osapuilleen standardinormaalijakauma $N(0,1)$, mutta tällä kertaa tätä tietoa käytetään hyväksi hienostuneemmalla tavalla. Nyt luottamusväli muodostetaan ratkaisemalla epäyhtälöpari

$$-z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{1}{n}\,p(1-p)}} \le z_{\alpha/2}$$

muuttujan $p$ suhteen toisen asteen yhtälön ratkaisukaavan avulla. Tuloksena saadaan Wilsonin luottamusväli

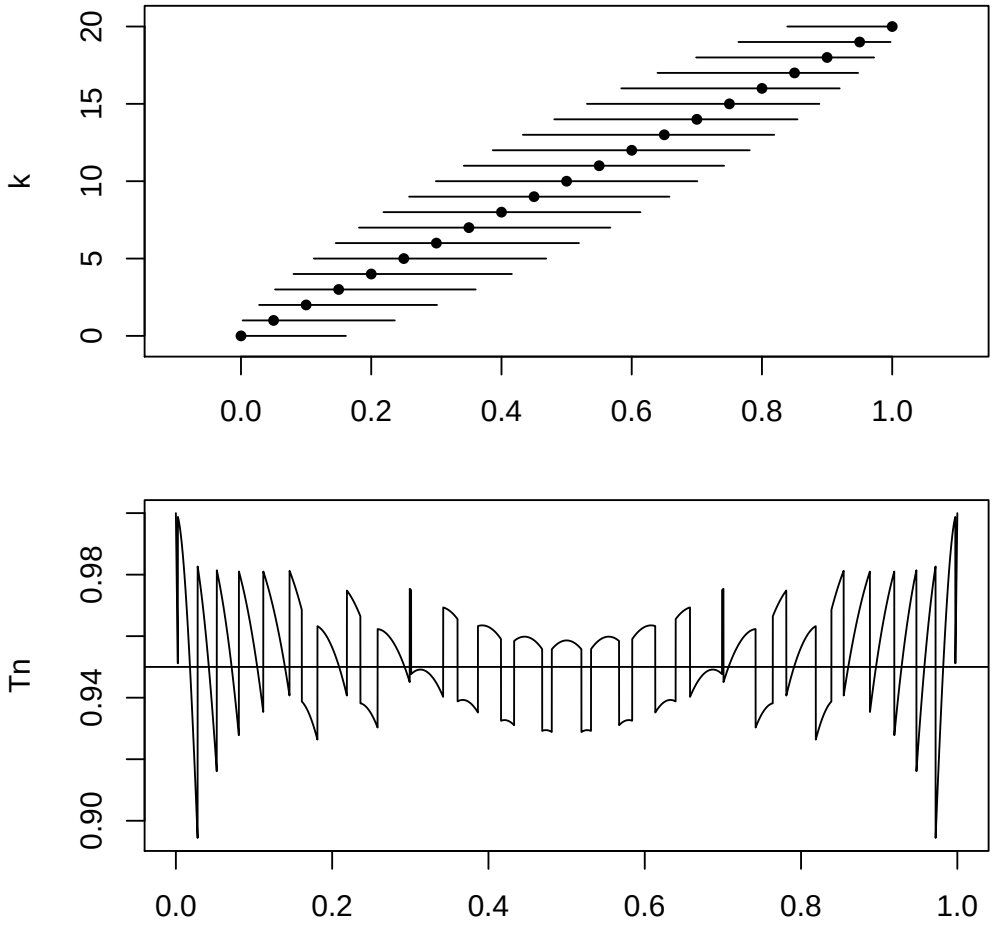
$$\frac{\hat{p} + \frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\,\hat{p}(1-\hat{p}) + \frac{1}{4n^2}\,z^2_{\alpha/2}}}{1 + \frac{1}{n}z^2_{\alpha/2}}, \tag{5.21}$$

joka on luottamusväliä (5.20) selkeästi parempi. (Luottamusväliä kutsutaan myös nimellä *Wilson score interval*, sen takia, että se voidaan johtaa invertoimalla tässä tilanteessa ns. pistemäärätesti, engl. *score test*.) Myös Wilsonin luottamusväli on likimääräinen, sillä luottamusvälin määritelmän epäyhtälö (5.2) ei sille toteudu. Kuvassa 5.8 esitetään Wilsonin luottamusvälin toiminta, kun $n = 20$. Tämä luottamusväli ei surkastu pisteeksi, jos onnistumisia on nolla tai $n$.

Clopper ja Pearson esittivät v. 1934 erään tavan muodostaa ns. tarkka (engl. *exact*) luottamusväli onnistumistodennäköisyydelle. Termi tarkka tarkoittaa tässä sitä, että kyseinen luottamusväli ei ole likimääräinen, vaan määritelmän (ks. kaava (5.2)) mukainen, eli

$$P_p(L(\mathbf{Y}) \le p \le U(\mathbf{Y})) \ge 1 - \alpha, \quad \text{kaikilla } 0 < p < 1.$$

**Kuva 5.8** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat Wilsonin luottamusvälit (5.21), kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys $p$:n funktiona. Nimellinen luottamustaso on osoitettu vaakaviivalla.



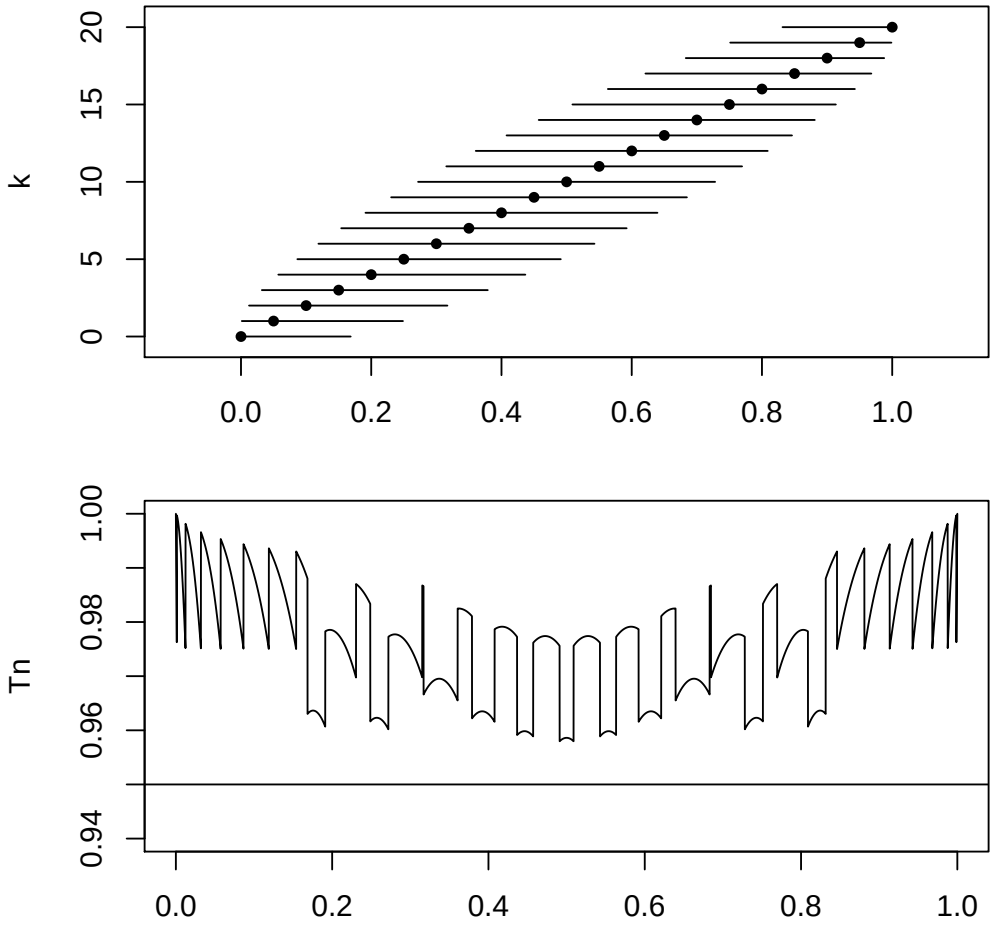
Lisäksi alarajaa $1-\alpha$ ei voida yhtään suurentaa ilman, että epäyhtälö rikkoontuisi jollakin otoskoolla $n$ ja jollakin $0 < p < 1$ . Muualla väli on turhan konservatiivinen, eli sen todellinen peittotodennäköisyys on aidosti lukua $1-\alpha$ suurempi, kuten kuvasta 5.9 nähdään, kun otoskoko $n = 20$.

Silloin kuin havaintosatunnaisvektorin jakauma on diskreetti, niin yleensä aina joudutaan tekemään luottamusvälien kanssa samantapaisia kompromisseja. Joko käytetään likimääräisiä luottamusvälejä, joiden todellinen peittotodennäköisyys on joskus pienempi kuin niiden nimellinen peittotodennäköisyys, tai sitten käytetään tarkkaa luottamusväliä (mikäli sellainen sattuu olemaan saatavilla), joka on useimmilla parametrinarvoilla turhan konservatiivinen.

Tietokoneella minkä tahansa edellä mainitun binomijakauman luottamusvälin laskeminen on yhtä helppoa. Esim. R-ohjelmistossa nämä luottamusvälit on helppo laskea `Hmisc`-kirjaston funktiolla `binconf`. Nimellistä luottamustasoa 95% vastaavat välit saadaan laskettua seuraavalla tavalla. (Myös funktio `binom.test` laskee Clopperin ja Pearsonin tarkan luottamusvälin. Funktio `prop.test` laskee erään luottamusvälin, joka on sukua Wilsonin luottamusvälille. Valitettavasti tämän funktion dokumentaatiosta on vaikea saada selvää.)

```
n <- 20
k <- 4
library(Hmisc)

## Loading required package: survival
## Loading required package: splines
## Hmisc library by Frank E Harrell Jr
##
## Type library(help='Hmisc'), ?Overview, or ?Hmisc.Overview')
## to see overall documentation.
##
## NOTE:Hmisc no longer redefines [.factor to drop unused levels when
## subsetting.  To get the old behavior of Hmisc type dropUnusedLevels().
##
## Attaching package: 'Hmisc'
## The following object(s) are masked from 'package:survival':
##
##    untangle.specials
## The following object(s) are masked from 'package:base':
##
##    format.pval, round.POSIXt, trunc.POSIXt, units

binconf(k, n, method = "asymptotic")

##  PointEst  Lower  Upper
##       0.2 0.0247 0.3753

binconf(k, n, method = "wilson")

##  PointEst   Lower Upper
##       0.2 0.08066 0.416
```

**Kuva 5.9** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat luottamustasoa 95% vastaavat Clopperin–Pearsonin tarkat luottamusvälit, kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys $p$:n funktiona.



```
binconf(k, n, method = "exact")

##  PointEst   Lower  Upper
##       0.2 0.05733 0.4366
```

## 5.9 Ennusteväli

Luottamusvälien lisäksi (tai sijasta) usein on mielekästä tarkastella aivan toisentyyppisiä välejä, ks. esim. Vardeman [3]. Käsittelemme tässä vain ennusteväliä. Vardeman esittelee myös ns. toleranssivälin.

Tarkastelemme yksinkertaisuuden vuoksi teoreettista populaatiota, jossa satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ ovat riippumattomia ja samoin jakautuneita satunnaismuuttujia pistetodennäköisyysfunktiolla tai tiheysfunktiolla $g(y;\theta)$. Välit pitää muodostaa $n$ ensimmäisen satunnaismuuttujan $Y_1, \ldots, Y_n$ arvojen avulla, ja tavalliseen tapaan,

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

Satunnaismuuttujan $Y_{n+1}$ ajatellaan olevan tulevaisuudessa saatava havainto tästä samasta jakaumasta.

**Määritelmä 5.6** (Ennusteväli)**.** Aineistosta laskettu väli $[L(\mathbf{y}), U(\mathbf{y})]$ on tason $1-\alpha$ *ennusteväli* (engl. *prediction interval*) satunnaismuuttujalle $Y_{n+1}$, jos vastaava satunnainen väli $[L(\mathbf{Y}), U(\mathbf{Y})]$ toteuttaa vaatimuksen

$$P_\theta\left(L(\mathbf{Y}) \leq Y_{n+1} \leq U(\mathbf{Y})\right) \geq 1-\alpha, \quad \text{kaikilla } \theta \in \Theta. \tag{5.22}$$

**Esimerkki 5.3** Jos normaalijakaumaa $N(\mu, \sigma^2)$ noudattavan populaation varianssi on tunnettu luku, ja $\bar{Y}$ on $n$ ensimmäisen satunnaismuuttujan otoskeskiarvo, niin

$$Y_{n+1} - \bar{Y} \sim N(0, (1+\frac{1}{n})\sigma^2)$$

Tästä nähdään helpoilla laskuilla, että todennäköisyydellä $1-\alpha$

$$Y_{n+1} \in \bar{Y} \pm z_{\alpha/2}\sqrt{1+\frac{1}{n}}\,\sigma$$

kaikilla $\mu$, joten tätä vastaava aineistosta laskettu väli on tason $1-\alpha$ ennusteväli.

Huomaa, että uuden havainnon ennusteväli on *paljon leveämpi* kuin odotusarvon $\mu$ kaksisuuntainen luottamusväli (5.10).

Jos myös varianssiparametri olisi tuntematon, niin ennusteväliä lähdetään konstruoimaan sillä perusteella, että

$$\begin{aligned} Y_{n+1} - \bar{Y} &\sim N(0, (1+\frac{1}{n})\sigma^2) \\ \frac{n-1}{\sigma^2}\,S^2 &\sim \chi^2_{n-1}, \end{aligned}$$

jossa otosvarianssi $S^2$ lasketaan satunnaismuuttujista $Y_1, \ldots, Y_n$. Yllä nämä kaksi satunnaismuuttujat ovat lisäksi riippumattomia. Tästä havainnosta saadaan yksinkertaisilla laskuilla aikaan ennusteväli uudelle havainnolle $Y_{n+1}$ käyttämällä $t$-jakauman kvantiileja. △

# Kirjallisuutta

[1] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–116, 2001.

[2] Robert G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17:857–872, 1998.

[3] Stephen B. Vardeman. What about the other intervals? *The American Statistician*, 46:193–197, 1992.

# Luku 6

# Tilastollinen testaus

Tilastolliset testit ovat frekventistisen päättelyn käytetyimpiä (ja valitettavasti myös huonoiten ymmärrettyjä ja sen takia myös eniten väärinkäytettyjä) tilastollisen päättelyn menetelmiä. Niiden avulla pyritään ottamaan kantaa tilastollisia malleja koskeviin väitteisiin eli hypoteeseihin, kuten esim.

- Onko tutkittava lantti harhaton?
- Onko tietyllä käsittelyllä vaikutusta? (Käsittely voisi olla esimerkiksi uusi hoitomuoto jollekin sairaudelle tai uusi lannoite tai uusi opetusmenetelmä.)

Testaus sujuu käytännössä laskemalla tilanteeseen sopivan testisuureen arvo. Laskettua arvoa verrataan siihen, minkälaisia arvoja hypoteesin mukaisesta populaatiosta satunnaisvaihtelu huomioon ottaen pitäisi tulla. Mikäli laskettu testisuureen arvo poikkeaa riittävän paljon hypoteesin mukaisista tyypillisistä arvoista, hypoteesi hylätään. Tällöin meistä ei enää ole uskottavaa, että näin suuri poikkeama aiheutuisi satunnaisvaihtelusta, vaan pidämme uskottavampana selityksenä sitä, että asetettu hypoteesi ei pidä paikkaansa.

## 6.1 Testauksen peruskäsitteitä

Tarkastelemme testausta frekventistisessä viitekehyksessä parametrisessa tilastollisessa mallissa eli jakaumaperheessä

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

Koska havaintoja vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma tunnetaan täysin, mikäli parametrinarvo $\theta$ tunnetaan, niin vektorin $\mathbf{Y}$ jakaumaa koskevat väittämät eli (tilastolliset) hypoteesit voidaan pukea siihen muotoon, että parametri kuuluu johonkin tiettyyn parametrialueen osajoukkoon.

Esimerkiksi lantinheittoa tavallisesti mallinnetetaan binomikokeena, jossa onnistumistodennäköisyys $\theta$ on tuntematon ja toistojen lukumäärä $n$ on tunnettu. Väite "lantti on harhaton" vastaa hypoteesia

$$\theta = \frac{1}{2}, \quad \text{eli} \quad \theta \in \{\frac{1}{2}\}.$$

Lantin harhattomuutta koskeva hypoteesi on *yksinkertainen* (engl. *simple*) eli *täysin määrätty*, sillä hypoteesia vastaava parametriavaruuden osajoukko

sisältää vain yhden pisteen $\theta_0 = \frac{1}{2}$. Tällöin satunnaisvektorin $\mathbf{Y}$ jakaumalla on yptnf/ytf $f(\mathbf{y}; \theta_0)$, mikäli hypoteesi on tosi. Paljon tyypillisempää on, että hypoteesi on *yhdistetty* (engl. *composite*) eli *osittain määrätty*, mikä tarkoittaa sitä, että hypoteesia vastaava parametriavaruuden osajoukko koostuu useammasta kuin yhdestä pisteestä.

Jotkin hypoteesit saattavat ensinäkemältä vaikuttaa yksinkertaisilta, vaikka ne todellisuudessa ovat yhdistettyjä. Jos esimerkiksi normaalijakautuneessa populaatiossa $N(\mu, \sigma^2)$ molemmat parametrit ovat tuntemattomia, niin tällöin parametria $\mu$ koskeva tarkka hypoteesi

$$H : \mu = 0$$

on yhdistetty hypoteesi, sillä se vastaa parametriavaruuden osajoukkoa

$$\{(\mu, \sigma^2) : \mu = 0 \text{ ja } \sigma^2 > 0\}.$$

Testauksessa asetetaan ns. *nollahypoteesi* (engl. *null hypothesis*) $H_0$, joka on muotoa

$$H_0 : \theta \in \Theta_0, \tag{6.1}$$

missä $\Theta_0 \subset \Theta$ on ei-tyhjä parametriavaruuden osajoukko. Testauksen tavoitteena on arvioida havaintojen $\mathbf{y}$ avulla nollahypoteesin paikkansapitävyyttä.

Tavallisesti nollahypoteesi vastaa vakiintunutta teoriaa tai sitä pessimististä näkemystä, että käsittelyllä ei ole vaikutusta. Tällöin tutkija tahtoisi todellisuudessa löytää todisteita nollahypoteesin hylkäämiseksi, mutta koska vakiintunutta teoriaa ei voida hylätä löyhin perustein, sitä vastaan pitää saada vakuuttavia todisteita, ennen kuin yhteisö suostuu hylkäämään nollahypoteesin.

Monesti nollahypoteesin $H_0$ lisäksi muotoillaan myös *vaihtoehtoinen hypoteesi* eli *vastahypoteesi* (engl. *alternative hypothesis*, myös *study hypothesis*), jota tyypillisesti merkitään symbolilla $H_1$ (tai toisinaan $H_A$). Vastahypoteesin mukaan $\theta$ kuuluu parametriavaruuden osajoukkoon $\Theta_1$, ts.

$$H_1 : \theta \in \Theta_1. \tag{6.2}$$

Tällöin vähintäänkin oletetaan, että

$$\Theta_0 \cap \Theta_1 = \emptyset,$$

ja usein (mutta ei aina) pätee $\Theta = \Theta_0 \cup \Theta_1$. Jos vastahypoteesi asetetaan, niin tämä tarkoittaa kannanottoa sen suhteen, mitä parametrin ajatellaan toteuttavan siinä tapauksessa, että nollahypoteesi osoittautuu epäilyksen alaiseksi.

Nollahypoteesia ja vastahypoteesia käsitellään testauksessa epäsymmetrisellä tavalla. Tämän asian ymmärtäminen on edellytys sille, että osaamme tulkita testin lopputuloksen järkevästi. Tilanteen voi ajatella olevan analoginen oikeudenkäynnin kanssa, jossa syytettynä on nollahypoteesi $H_0$. $H_0$ on oikeudenkäynnissä syytön, ellei sitä osoiteta syylliseksi.

Testaus sujuu siten, että aineistosta lasketaan tunnusluku $t(\mathbf{y})$, jota kutsutaan testisuureeksi (engl. *test statistic*). Testisuure asettaa mahdolliset havainnot järjestykseen jollakin seuraavista tavoista sen asian suhteen, miten tyypillisinä tai outoina niitä pidämme, jos nollahypoteesi pitää paikkansa.

1. Pienet tunnusluvun arvot viittaavat siihen, että aineisto on sopusoinnussa $H_0$:n kanssa, ja suuret viittaavat ristiriitaan $H_0$:n kanssa.

2. Suuret tunnusluvun arvot viittavat sopusointuun $H_0$:n kanssa ja pienet arvot ristiriitaan sen kanssa.

3. Suuri poikkeama jostakin vertailuarvosta $t_0$ ylöspäin tai alaspäin viittaa ristiriitaan ja pieni poikkeama sopusointuun.

Lisäksi vaaditaan se, että hallitsemme testisuuretta vastaavan satunnaismuuttujan $t(\mathbf{Y})$ jakauman ainakin kaikilla nollahypoteesin mukaisilla parametrinarvoilla $\theta \in \Theta_0$. Usein testisuureena käytetään saranasuuretta, mikäli sellainen tunnetaan. Tällä kurssilla emme voi paneutua tämän syvällisemmin siihen, kuinka testisuure pitäisi valita.

Tilastollinen testi toimii sillä tavalla, että havaitusta aineistosta lasketaan testisuureen arvo, ja sitten tarkistetaan kuuluuko se *kriittiseen alueeseen* (engl. *critical region*) eli *hylkäysalueeseen* (engl. *rejection region*) $C$. Testi antaa yhden kahdesta vaihtoehtoisesta päätöksestä: se joko hylkää nollahypoteesin tai sitten ei sen mukaan, kuuluko testisuureen arvo kriittiseen alueeseen vai ei.

- Jos $t(\mathbf{y}) \in C$, niin testi *hylkää* (engl. *reject*) nollahypoteesin, eli testi on *merkitsevä* (engl. *significant*).

- Jos $t(\mathbf{y}) \notin C$, niin testi *hyväksyy* (engl. *accept*) nollahypoteesin (mikä voidaan ilmaista myös sanomalla, että *nollahypoteesi jää voimaan*), eli testi *ei ole merkitsevä* (engl. *not significant*).

Huomaa, että hylkääminen ja sen vastakohta, jota yksinkertaisuuden vuoksi tavallisimmin kutsutaan hyväksymiseksi, ovat testaukseen liittyviä teknisiä termejä. Se mitä käytännön johtopäätöksiä ja käytännön toimia testin lopputuloksen selvittyä tehdään, on eri asia kuin testin antama päätös. Varsinkin termi hyväksyä on harhaanjohtava. Mikäli $H_0$ hyväksytään, niin tutkija usein oikeasti edelleen epäilee nollahypoteesin paikkansapitävyyttä, mutta hän ei ole löytänyt aineistosta riittävän vakuttavaa todistetta sitä vastaan.

Kriittisen alueen muoto riippuu siitä, minkälaiset tunnusluvun arvot ovat nollahypoteesin kanssa yhteensopimattomia. Jos suuret tunnusluvun arvot ovat nollahypoteesin kannalta kriittisiä, niin kriittinen alue on muotoa

$$C = (u, \infty)$$

ts. testi hylkää nollahypoteesin, jos $t(\mathbf{y}) > u$. Tällöin kynnysarvoa $u$ voidaan kutsua *kriittiseksi arvoksi* (engl. *critical value*). Tavallisesti kriittinen alue määräytyy testin merkitsevyystasosta.

Testien yhteydessä puhutaan niiden koosta tai merkitsevyystasosta. Käytämme näitä termejä synonyymeinä, mutta jotkut kirjoittajat tekevät näiden käsitteiden välille eron.

**Määritelmä 6.1.** Jos testin kriittinen alue on $C$, niin testin *koko* (engl. *size*) eli sen *merkitsevyystaso* (engl. *significance level*) on

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(t(\mathbf{Y}) \in C) = \sup_{\theta \in \Theta_0} P_\theta(H_0 \text{ hylätään}) \tag{6.3}$$

Tässä sup eli *supremum* tarkoittaa pienintä ylärajaa; testin koko $\alpha$ on pienin yläraja hylkäystodennäköisyydelle $P_\theta(t(\mathbf{Y}) \in C)$, kun satunnaisvektorilla $\mathbf{Y}$ on nollahypoteesin mukainen jakauma. Ts. $0 < \alpha < 1$, ja

$$P_\theta(t(\mathbf{Y}) \in C) \leq \alpha, \qquad \text{kaikilla } \theta \in \Theta_0$$

ja kaikilla $\epsilon > 0$ on olemassa $\theta \in \Theta_0$ siten, että

$$P_\theta(t(\mathbf{Y}) \in C) > \alpha - \epsilon.$$

Usein hylkäystodennäköisyys $P_\theta(t(\mathbf{Y}) \in C)$ pysyy vakiona joukossa $\Theta_0$. Näin käy automaattisesti, jos $H_0$ on yksinkertainen hypoteesi ja myös silloin, jos testisuure on saranasuure. Tässä tapauksessa testin merkitsevyystaso on yksinkertaisesti

$$\alpha = P_\theta(H_0 \text{ hylätään}), \qquad \text{millä tahansa } \theta \in \Theta_0. \tag{6.4}$$

Tyypillisesti testin merkitsevyystaso $0 < \alpha < 1$ asetetaan, ja sitten tämän informaation perusteella määritetään kriittinen alue $C$ siten, että vaatimus (6.3) toteutuu. Ennen vanhaan ei ollut käytössä tilastollisia ohjelmia, ja merkitsevyystasolle $\alpha$ kiinnitettiin tavallisesti jokin seuraavista konventionaalisista arvoista

$$0.05, \quad 0.01, \quad \text{tai} \quad 0.001$$

sen takia, että näitä arvoja vastaavat kriittiset arvot löytyivät tilastollisista taulukoista. Nämä konventionaaliset tasot ovat mielivaltaisia, ja ne on valittu sillä perusteella, että vastaavat murtoluvut (yksi kahdestakymmenestä, yksi sadasta, yksi tuhannesta) ovat pyöreitä.

Testin tekemään päätökseen liittyy aina virheen mahdollisuus. Jos $H_0$ pitää paikkansa, mutta testi hylkää sen, tällöin tapahtuu *hylkäämisvirhe* eli *I lajin virhe* (engl. *type I error*). Jos $H_1$ pitää paikkansa, mutta testi hyväksyy $H_0$:n, tapahtuu *hyväksymisvirhe* eli *II lajin virhe* (engl. *type II error*).

| Todellisuus | Päätös | |
|---|---|---|
| | $H_0$ hyväksytään | $H_0$ hylätään |
| $H_0$ tosi | oikea päätös | hylkäämisvirhe<br>I lajin virhe |
| $H_1$ tosi | hyväksymisvirhe<br>II lajin virhe | oikea päätös |

Testissä nollahypoteesia ja vastahypoteesia kohdellaan epäsymmetrisellä tavalla. Merkitsevyystaso $\alpha$ on yläraja hylkäämisvirheen todennäköisyydelle. *Jos nollahypoteesi pitää paikkaansa*, niin testisuure saa hylkäämiseen johtavia arvoja niin harvoin, että hylkäämistodennäköisyys on enintään $\alpha$. Tähän asti emme ole lainkaan miettineet sitä, mitä testissä tapahtuu jos $H_1$ on tosi.

Perinteinen tapa raportoida testin tulos on ollut kiinnittää testin koko $\alpha$ sekä kertoa testin päätös, eli hylkäsikö vai hyväksyikö testi nollahypoteesin, mutta nykyään ei välttämättä toimita näin yksioikoisesti.

## 6.2 Normaalijakautuneen populaation odotusarvon testaus, kun varianssi on tunnettu

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Ts. satunnaismuuttujat $Y_i$ ovat riippumattomia, ja niillä on kaikilla normaalijakauma $N(\mu, \sigma^2)$.

Tarkastelemme odotusarvon testausta, kun varianssi on tunnettu. Tällä tilanteella ei ole suurta käytännön arvoa. Tästä syystä esim. R-ohjelmistossa ei ole valmista funktiota, joka laskisi kätevästi tämän testin tulokset. Tätä testiä käsitellään sen vuoksi, että teoria on tässä tapauksessa helppoa.

**Yksisuuntainen testi**

Jos populaation varianssi $\sigma^2$ on tunnettu luku, niin

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

on saranasuure. Tarkastelemme nollahypoteesia

$$H_0 : \mu = \mu_0,$$

jossa $\mu_0$ on tunnettu luku (esim. $\mu_0 = 0$). Tämä on yksinkertainen hypoteesi. Otamme ensin vastahypoteesiksi yksisuuntaisen hypoteesin

$$H_1 : \mu > \mu_0,$$

joka on yhdistetty hypoteesi. Tätä hypoteesiparia vastaavat parametriavaruuden osajoukot

$$\Theta_0 = \{\mu_0\}, \qquad \Theta_1 = (\mu_0, \infty)$$

Käytämme testisuureena tunnuslukua

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}. \tag{6.5}$$

Huomaa, että testisuure on tunnusluku (toisin kuin sitä vastaava saranasuure), sillä testisuureessa tuntemattoman parametrin $\mu$ tilalla on tunnettu arvo $\mu_0$. Testisuuretta vastaavalla satunnaismuuttujalla

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

on standardinormaalijakauma $N(0,1)$, kun nollahypoteesi pitää paikkansa. Nyt suuret testisuureen arvot ovat nollahypoteesin kannalta kriittisiä, sillä $\bar{Y}$ estimoi populaatioparametria $\mu$, ja testisuure on kasvava funktio tästä estimaattorista.

Tason $\alpha$ testi saadaan aikaan käyttämällä kriittistä arvoa $z_\alpha$, sillä

$$P_{\mu_0}(t(\mathbf{Y}) > z_\alpha) = P(Z > z_\alpha) = \alpha,$$

jossa $Z \sim N(0,1)$. Tästä nähdään, että luottamustason $\alpha$ testi hypoteesiparille

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0,$$

tekee päätöksen seuraavasti. Ensin lasketaan testisuureen arvo

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

ja sitten testi toimii seuraavasti

$$\begin{cases} \text{jos } z > z_\alpha, & H_0 \text{ hylätään} \\ \text{jos } z \leq z_\alpha & H_0 \text{ hyväksytään.} \end{cases}$$

Ts. testin hylkää nollahypoteesin silloin (ja vain silloin), kun

$$z > z_\alpha. \tag{6.6}$$

Tämä on ns. *yksisuuntainen* eli *yksitahoinen z-testi* (engl. *one-sided* tai *one-tailed z-test)*.

Tällä tavalla muotoiltuna yksisuuntainen testi on omituinen, sillä

$$\Theta_0 \cup \Theta_1 \neq \Theta,$$

vaan parametriavaruudesta jätetään kokonaan huomioimatta ne $\mu$, joille $\mu < \mu_0$. On vaikea sanoa esim., tehdäänkö virhe vai toimitaanko oikein, jos todellisuudessa $\mu < \mu_0$, mutta nollahypoteesi hylätään.

Näemme myöhemmin, että sama yksisuuntainen testi on koon $\alpha$ testi myös hypoteesiparille

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0.$$

Tälle hypoteesiparille

$$\Theta_0 \cup \Theta_1 = (-\infty, \mu_0] \cup (\mu_0, \infty) = \mathbb{R} = \Theta.$$

On järkevää ajatella, että yksisuuntaisella testillä (6.6) selvitetään tämän jälkimmäisen yhdistetyn nollahypoteesin $\mu \leq \mu_0$ paikkansapitävyttä. Tämän testin kriittinen alue sattuu olemaan paljon helpompi johtaa, jos nollahypoteesina käytetään yksinkertaista hypoteesia $\mu = \mu_0$. Tämä lienee se ainoa syy, miksi tätä tarkkaa nollahypoteesin muotoilua lainkaan käytetään yksisuuntaiselle $z$-testille.

Vastaavilla laskuilla nähdään, että sekä hypoteesiparille

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu < \mu_0$$

että hypoteesiparille

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

luottamustason $\alpha$ testi tekee päätöksen seuraavasti. Ensin lasketaan testisuureen arvo

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}.$$

Tällä kertaa pienet arvot (ts. itseisarvoltaan suureet negatiiviset arvot) ovat nollahypoteesille kriittisiä. Testi hylkää nollahypoteesin silloin (ja vain silloin), kun

$$z < -z_\alpha \tag{6.7}$$

**Kaksisuuntainen testi**

Nyt nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

Testisuure on edelleen

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

ja sitä vastaavalla satunnaismuuttujalla $t(\mathbf{Y})$ on $N(0,1)$-jakauma, kun $H_0$ pitää paikkansa. Nyt sekä suuret että pienet testisuureen arvot ovat nollahypoteesin

kannalta kriittisiä. Kaksisuuntainen (engl. *two-sided*, *two-tailed*) $z$-testi luottamustasolla $\alpha$ hylkää nollahypoteesin (täsmälleen) silloin, kun

$$|z| > z_{\alpha/2}. \tag{6.8}$$

Tämä perustuu siihen, että

$$P(|Z| > z_{\alpha/2}) = \alpha,$$

kun $Z \sim N(0,1)$.

**Fiktiivinen numeerinen esimerkki**

Planeetalla Z seurataan tiiviisti JTP-kurssin aineistoja, koska paikalliset tutkijat ovat huomanneet kosmisen yhteyden planeetan Z ilmaston tilan ja JTP-kurssin simuloitujen aineistojen parametrien, erityisesti kuvan 4.3 aineiston parametrien välillä. Valitettavasti planeetalla Z ei osata suomea, vaan koko väestö puhuu englantia (hassusti murtaen). Tämän takia kukaan tutkija ei ole saanut selville, että oikeasti $\mu = 0.2012$. Sen sijaan ollaan saatu selville, että kyseessä on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, jossa $\sigma^2 = 1$. Planeetalla Z suositaan $z$-testejä merkitsevyystasolla 0.05.

Pessimistisimmät tutkijat ovat sitä mieltä, että $\mu \leq 0$, mikä tarkoittaa käytännössä koko planeetan pikaista tuhoa. Tämän takia teemme $z$-testin, jossa hypoteesit ovat

$$H_0 : \mu \leq 0, \qquad H_1 : \mu > 0.$$

Aineistossa

$$\bar{y} = 0.726, \quad n = 10,$$

Nollahypoteesia vastaava, aineistosta laskettu $z$-arvo on

$$z = \frac{\bar{y} - 0}{\sigma/\sqrt{n}} = 2.296$$

Koska $z > z_{0.05} = 1.645$, *nollahypoteesi* $\mu \leq 0$ *hylätään* merkitsevyystasolla 5 %. Planeetan sanomalehdet kirjoittavat etusivuillaan, että tutkijat ovat todistaneet, että maailmanloppua ei tule.

Suurin osa tutkijoista ei kuitenkaan hyväksy hypoteesin $\mu \leq 0$ tieteellistä relevanssia, vaan uskoo teoriaan, jonka mukaan $\mu = \frac{1}{2}$, joka tarkoittaa sitä että planeetan ilmasto säilyy ikuisesti yhtä suotuisana kuin nykyään. Tämän takia teemme $z$-testin, jossa hypoteesit ovat

$$H_0 : \mu = \frac{1}{2}, \qquad H_1 : \mu \neq \frac{1}{2}.$$

Tätä nollahypoteesia vastaava aineistosta laskettu $z$-arvo on

$$z = \frac{\bar{y} - \frac{1}{2}}{\sigma/\sqrt{n}} = 0.715,$$

Koska $|z| \leq z_{\alpha/2} = 1.960$, niin *nollahypoteesi* $\mu = \frac{1}{2}$ *hyväksytään* merkitsevyystasolla 5 %. Planeetan sanomalehdet kirjoittavat etusivullaan, että tutkijat ovat todistaneet, että ilmasto säilyy ikuisesti suotuisana.

Mikä tulosten uutisoinnissa oli vikana? Nollahypoteesin hylkääminen tarkoittaa sitä, että ollaan löydetty todisteita sitä vastaan. Nollahypoteesin hyväksyminen ei tarkoita sitä, että oltaisiin löydetty todisteita nollahypoteesin puolesta. Se tarkoittaa sitä, että ei olla löydetty painavia todisteita nollahypoteesia vastaan.

Mitä varten tässä esimerkissä hölmö ja epätosi nollahypoteesi $H_0 : \mu = \frac{1}{2}$ hyväksyttiin?

Tähän asiaan saamme lisävalaistusta sen jälkeen, kun olemme nähneet, kuinka $z$-testin voima saadaan laskettua.

## 6.3 Testin voima

**Määritelmä 6.2** (Testin voima)**.** Jos $C$ on tarkasteltavan testin kriittinen alue, niin parametriavaruudella määriteltyä funktio

$$\pi(\theta) = P_\theta(t(\mathbf{Y}) \in C) = P_\theta(H_0 \text{ hylätään}) \tag{6.9}$$

on nimeltään testin *voima* (engl. *power*) tai sen voimafunktio (engl. *power function*).

Toisin sanoen, voimafunktio on testin hylkäystodennäköisyys parametrin funktiona. (Se mittaa *hylkäys*voimaa.)

Testin voiman voi ilmaista monimutkaisemmalla tavalla ensimmäisen ja toisen lajin virheiden todennäköisyyden avulla, nimittäin

$$\pi(\theta) = \begin{cases} P_\theta(\text{I lajin virhe}), & \text{kun } \theta \in \Theta_0, \\ 1 - P_\theta(\text{II lajin virhe}), & \text{kun } \theta \in \Theta_1. \end{cases}$$

Jos testin koko on $\alpha$, niin testin koon määritelmän (6.3) ja sen voiman määritelmän (6.9) mukaan testin koko (ts. sen merkitsevyystaso) on

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta).$$

Merkitsevyystaso asettaa siis ylärajan testin voimalle nollahypoteesin mukaisilla parametrinarvoilla. Luonnollisesti tahtoisimme, että testin voima olisi mahdollisimman suuri vaihtoehtohypoteesin mukaisilla parametrinarvoilla.

Testin voimaa on syytä tarkastella tutkimuksen suunnitteluvaiheessa. Yleisesti ottaen, *mitä suurempi on otoskoko, sitä suurempi on testin voima* (vaihtoehtohypoteesin mukaisilla parametrinarvoilla). Tavallisesti tutkijalla on käsitys siitä, miten suuret poikkeamat nollahypoteesin mukaisista parametrinarvoista ovat käytännössä merkittäviä. Tällöin otoskoko voidaan yrittää valita siten, että saavutetaan vähintään jokin annettu voima (esim. vähintään 80 %) aina, kun poikkeama nollahypoteesista on käytännön kannalta merkittävä.

Tietyt tieteelliset lehdet vaativat laskelmaa tilastollisen testin voimasta, mikäli sellaisia artikkelissa esitetään. Esim. *American Psychological Association* -yhdistyksen lehtien julkaisuohjeissa (6. laitos, v. 2010) ohjeistetaan

> routinely provide evidence that the study has sufficient power to detect effects of substantive interest.

## 6.4 Testin $p$-arvo

Vanhanaikainen tapa raportoida testin tulos on kertoa testin koko $\alpha$ sekä kertoa testin päätös, eli hylkäsikö vai hyväksyikö testi nollahypoteesin. Nykyään on tapana kertoa tämän lisäksi (tai sijasta) testin *p-arvo* (engl. *p-value*) eli *havaittu merkitsevyystaso* (engl. *observed significance level*). Testin $p$-arvo mittaa tietyllä tavalla nollahypoteesin ja aineiston yhteensopivuutta.

Testin $p$-arvon määrittely on hieman erilainen sen mukaan, mitkä testisuureen arvot ovat nollahypoteesille kriittisiä. Mikäli testisuureen $t(\mathbf{y})$ suuret arvot ovat nollahypoteesille kriittisiä, niin $p$-arvo määritellään kaavalla

$$p = p(\mathbf{y}) = \sup_{\theta \in \Theta_0} P_\theta[t(\mathbf{Y}) \geq t(\mathbf{y})] \tag{6.10}$$

Tässä $t(\mathbf{y})$ on havaitusta aineistosta $\mathbf{y}$ laskettu testisuureen arvo, ja $t(\mathbf{Y})$ on satunnaisvektorista $\mathbf{Y}$ laskettu testisuureen arvo, kun sillä on jakaumana mallin mukainen yptnf/ytf $f(\mathbf{y};\theta)$. Useissa tilanteissa tässä merkitty todennäköisyys ei riipu siitä, mitä nollahypoteesin mukaista parametrinarvoa $\theta \in \Theta_0$ tarkastellaan, jolloin $p$-arvo voidaan määritellä sanallisesti seuraavasti.

> $p$-arvo on se todennäköisyys, jolla nollahypoteesin mukaisesta populaatiosta saadaan testisuureen arvo, joka on vähintään yhtä kummallinen kuin aineistosta laskettu testisuureen arvo.

Kummallisuutta mitataan testisuureen arvolla: kummallisempia ovat ne arvot jotka poikkeavat vielä enemmän nollahypoteesille kriittiseen suuntaan kuin havaittu arvo. Usein kummallisuuden sijasta sanotaan "vähintään yhtä äärevä" (engl. *at least as extreme as*).

Jos $p$-arvo on pieni (esim. 1 %), niin tällöin hyvin pienellä todennäköisyydellä nollahypoteesin mukaisesta populaatiosta saadaan vähintään yhtä kummallisia havaintoja kuin mitä todellisuudessa saatiin. Jos taas $p$-arvo on suuri (esim. 20 %), niin tällöin kohtuullisen usein nollahypoteesin mukaisesta populaatiosta saadaan havaintoja, jotka ovat vähintään yhtä kummallisia kuin mitä todellisuudessa havaittiin. Toisin sanoen

> pieni $p$-arvo viittaa ristiriitaan aineiston ja nollahypoteesin välillä.

Jos $p$-arvon määritelmässä (6.10) oleva todennäköisyys kuitenkin riippuu parametrinarvosta $\theta \in \Theta_0$, niin oikea sanallinen määritelmä on

- $p$-arvo on pienin yläraja sille todennäköisyydelle, jolla nollahypoteesin mukaisesta populaatiosta saadaan testisuureen arvo, joka on vähintään yhtä kummallinen kuin aineistosta laskettu testisuureen arvo.

Testin päätös saadaan luettua sen $p$-arvosta seuraavalla tavalla.

> Testi hylkää $H_0$:n, jos $p < \alpha$. Muussa tapauksessa $H_0$ jää voimaan.

Osoitamme esimerkeissä, että näin menetellen testin kooksi tulee $\alpha$.

Nykyään on tapana ilmoittaa testin $p$-arvo (parilla desimaalilla), ja lisäksi varmuuden vuoksi kommentoida (asiantuntematonta lukijaa ajatellen), tulisko nollahypoteesi hylättyä vai hyväksyttyä konventionaalisilla merkitsevyystasoilla. Tämä on paljon informatiivisempaa kuin kertoa vain testin päätös jollakin kiinteällä merkitsevyystasolla.

## 6.5 $z$-testin $p$-arvo ja voima

Laskemme seuraavaksi jaksossa 6.2 käsitellyn $z$-testin $p$-arvon ja voimafunktion sekä yksi- että kaksisuuntaisessa tapauksessa.

### Yksisuuntainen $z$-testi

Yksisuuntaisessa testissä

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0,$$

jossa testisuure lasketaan kaavalla

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

ovat suuret testisuureen arvot kriittisiä, ja $t(\mathbf{Y}) \sim N(0,1)$, kun nollahypoteesi pitää paikkansa. Tämän takia testin $p$-arvo on

$$p = P_{\mu_0}(Z \geq z) = 1 - \Phi(z),$$

jossa $\Phi$ on $N(0,1)$-jakauman kertymäfunktio, $Z \sim N(0,1)$, ja $z$ on aineistosta laskettu testisuureen arvo.

Tämä testi hylkää täsmälleen silloin, kun

$$\begin{aligned} & z > z_\alpha \\ \Leftrightarrow \quad & \Phi(z) > 1 - \alpha \\ \Leftrightarrow \quad & p = 1 - \Phi(z) < \alpha. \end{aligned}$$

Tällä tavalla olemme tarkistaneet yksisuuntaisen $z$-testin kohdalla, että testin päätös voidaan lukea sen $p$-arvosta.

Voimafunktio on

$$\pi(\mu) = P_\mu[t(\mathbf{Y}) > z_\alpha] = P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha\right],$$

ja tässä satunnaismuuttujan $(\bar{Y}-\mu_0)/(\sigma/\sqrt{n})$ jakauma on $N((\mu-\mu_0)/(\sigma/\sqrt{n}), 1)$, kun todellinen parametrinarvo on $\mu$, vrt. kuva 6.1. Siis

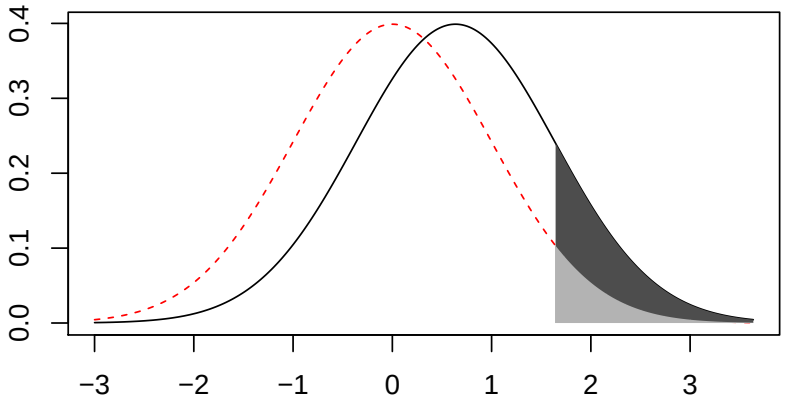
$$\begin{aligned} \pi(\mu) &= P_\mu\left[\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= P\left[Z > z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_\alpha\right). \end{aligned}$$

Viimeinen yhtäsuuruus perustuu $N(0,1)$-jakauman symmetrisyyteen, minkä takia sen kertymäfunktio toteuttaa identiteetin

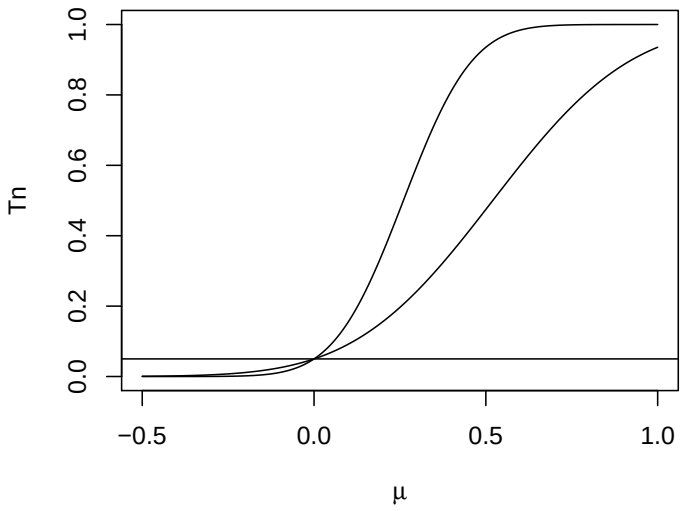
$$1 - \Phi(a) = \Phi(-a), \qquad \text{kaikilla } a.$$

Kuvassa 6.2 näytetään tämän testin voimafunktio, kun testin koko on $\alpha = 0.05$. Vaihtoehtohypoteesin mukaisilla parametrinarvoilla suuremmalla otoskoolla saavutetaan suurempi voima.

**Kuva 6.1** Hypoteesin $H_0 : \mu = 0$ voiman laskeminen, kun todellisuudessa $\mu = 0.2012$. Normaalijakauman varianssi $\sigma^2 = 1$ on oletettu tunnetuksi. Normaalijakauman $N(0, 1)$ häntäalueen pinta-ala on $\alpha = 0.05$, ja testin voima on normaalijakauman $N((\mu - 0)/(\sigma/\sqrt{n}), 1)$ kuvaan merkityn häntäalueen pinta-ala.




**Kuva 6.2** Yksisuuntaisen $z$-testin voimafunktio, kun $\alpha = 0.05$, $\mu_0 = 0$, $\sigma = 1$, ja otoskoko on $n = 10$ tai $n = 40$. Suuremmalla otoskoolla saavutetaan suurempi voima vastahypoteesin $\mu > \mu_0$ mukaisilla parametrinarvoilla. Testin koko on osoitettu vaakaviivalla.



Äsken johdetusta voimafunktion kaavasta (sekä kuvasta) nähdään, että se on aidosti kasvava funktio, minkä takia

$$\pi(\mu) \leq \pi(\mu_0) = \alpha, \quad \text{kaikilla } \mu \leq \mu_0.$$

Tämän takia käsittelemämme yksisuuntainen $z$-testi on tason $\alpha$ testi myös hypoteesiparille

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0.$$

Tälle hypoteesiparille

$$\Theta_0 \cup \Theta_1 = (-\infty, \mu_0] \cup (\mu_0, \infty) = \mathbb{R} = \Theta,$$

kuten aikaisemmin jo mainittiin.

Toisen mahdollisen yksisuuntaisen testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu < \mu_0$$

$p$-arvo on

$$p = P_{\mu_0}(Z \leq z) = \Phi(z),$$

jossa $\Phi$ on $N(0,1)$-jakauman kertymäfunktio, $Z \sim N(0,1)$, ja $z$ on aineistosta laskettu testisuureen arvo, eli

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

jonka pienet arvot ovat nollahypoteesille kriittisiä.

Tämän testin voimafunktio on

$$\begin{aligned}
\pi(\mu) &= P_\mu[t(\mathbf{Y}) \leq -z_\alpha] \\
&= P\left[Z \leq -z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\
&= \Phi\left(-z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Tämä funktio on aidosti vähenevä, joten

$$\pi(\mu) \leq \pi(\mu_0) = \alpha, \quad \text{kaikilla } \mu \geq \mu_0.$$

Voimafunktion kuvaaja on peilikuva ensimmäiseksi käsitellyn yksisuuntaisen $z$-testin voimafunktion kuvaajasta. Tämä yksisuuntainen $z$-testi on tason $\alpha$ testi myös hypoteesiparille

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

## Kaksisuuntainen $z$-testi

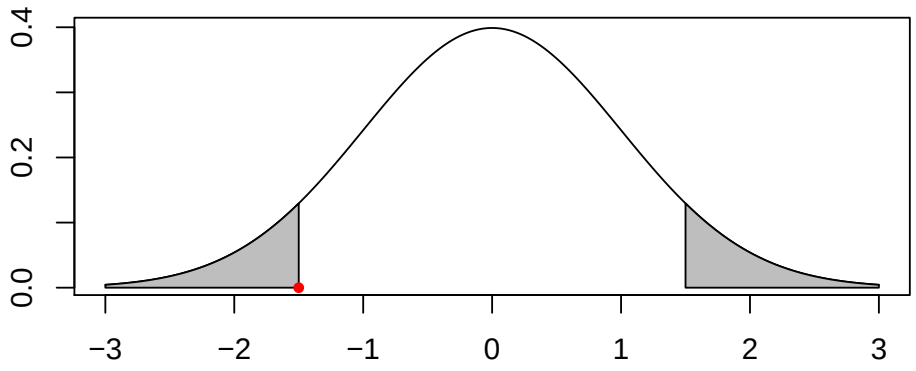
Kaksisuuntaisen testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

testaamisessa testisuureen

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

**Kuva 6.3** Kaksisuuntaisen testin $p$-arvon määrittäminen, kun havainnosta saadaan $z = -1.5$.




sekä suuret että pienet (negatiiviset) arvot ovat nollahypoteesille kriittisiä.

$P$-arvo määritellään summaamalla häntätodennäköisyys molemmilta viitejakauman hänniltä, ts. kummallisia tai vielä kummallisempi arvoja ovat ne, joille

$$|Z| \geq |z|,$$

ks. kuva 6.3. Tällä perusteella

$$p = P[|Z| \geq |z|] = \Phi(-|z|) + 1 - \Phi(|z|) = 2(1 - \Phi(|z|))$$

Kaksisuuntainen $z$-testi hylkää täsmälleen silloin, kun

$$\begin{aligned} & |z| > z_{\alpha/2} \\ \Leftrightarrow \quad & \Phi(|z|) > 1 - \frac{\alpha}{2} \\ \Leftrightarrow \quad & p = 2(1 - \Phi(|z|)) < \alpha. \end{aligned}$$

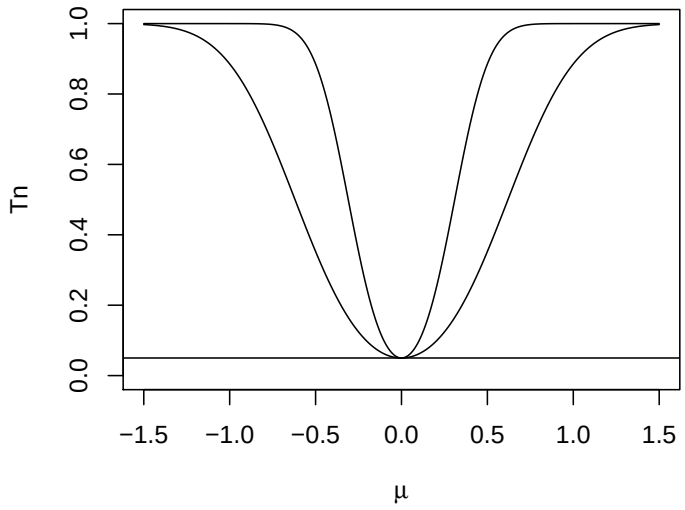
Nyt olemme tarkistaneet myös kaksisuuntiselle $z$-testille, että testin päätös voidaan lukea sen $p$-arvosta.

Kaksisuuntaisen testin voimafunktio on

$$\begin{aligned} \pi(\mu) &= P_\mu[|t(\mathbf{Y})| \geq z_{\alpha/2}] \\ &= P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}\right] + P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2}\right] \\ &= P\left[Z \geq z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] + P\left[Z \leq -z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= 1 - \Phi\left(z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Kuvassa 6.4 näytetään kaksisuuntaisen $z$-testin voimafunktio, kun testin koko on $\alpha = 0.05$. Suuremmalla otoskoolla saavutetaan suurempi voima.

**Kuva 6.4** Kaksisuuntaisen $z$-testin voimafunktio, kun $\alpha = 0.05$, $\mu_0 = 0$, $\sigma = 1$, ja otoskoko on $n = 10$ tai $n = 40$. Suuremmalla otoskoolla saavutetaan suurempi voima vastahypoteesin $\mu \neq \mu_0$ mukaisilla parametrinarvoilla. Testin koko on osoitettu vaakaviivalla.




## $p$-arvo ja voima numeerisessa esimerkissä

Palaamme planeetalle Z. Aineiston yhteenveto oli

$$\bar{y} = 0.726, \quad n = 10, \quad \sigma^2 = 1.$$

Aluksi testattiin yksisuuntaista hypoteesia

$$H_0 : \mu \leq 0, \qquad H_1 : \mu > 0,$$

joka hylättiin merkitsevyystasolla $\alpha = 0.05$. Tämän testin $p$-arvo ja testin voima todelliselle parametrinarvolle $\mu = 0.2012$ saadaan laskettua R:llä seuraavasti. Tässä funktio `pnorm` laskee normaalijakauman kertymäfunktion arvon.

```
mean.y <- 0.726
n <- 10
sigma <- 1
mu0 <- 0
alpha <- 0.05
mu.true <- 0.2012
sem <- sigma/sqrt(n)
z <- (mean.y - mu0)/sem
p <- 1 - pnorm(z)
zcrit <- qnorm(lower = FALSE, alpha)
pwr <- pnorm((mu.true - mu0)/sem - zcrit)
c(z, zcrit, p, pwr)

## [1] 2.29581 1.64485 0.01084 0.15658
```

Tulostuksista näemme, että testin $p$-arvo $p = 0.011$, joten tämä on yläraja sille todennäköisyydelle, että nollahypoteesin mukaisesta populaatiosta saadaan $z$-tunnusluvun arvo, joka on suurempi tai yhtä suuri kuin aineistosta laskettu $z$-tunnusluvun arvo. Testin voima todelliselle parametrinarvolle on vain 0.16. Ennen aineiston simulointia todennäköisyys, että testi tulee hylkäämään nollahypoteesin $\mu \leq 0$ oli (ainoastaan) 0.16, joten olimme aika onnekkaita koska pystyimme (todellisen tilanteen mukaisesti) hylkäämään tämän nollahypoteesin.

Toisen tutkijayhteisön mielipiteen mukaisesti seuraavaksi tehtiin testi

$$H_0 : \mu = \frac{1}{2}, \qquad H_1 : \mu \neq \frac{1}{2},$$

joka hyväksyttiin merkitsevyystasolla $\alpha = 0.05$. Tässä tapauksessa $p$-arvo ja testin voima voidaan laskea seuraavasti.

```
mu0 <- 0.5
z <- (mean.y - mu0)/sem
zcrit <- qnorm(alpha/2, lower = FALSE)
p <- 2 * (1 - pnorm(abs(z)))
pwr <- 1 - pnorm(zcrit - (mu.true - mu0)/sem) + pnorm(-zcrit -
    (mu.true - mu0)/sem)
c(z, zcrit, p, pwr)
```

```
## [1] 0.7147 1.9600 0.4748 0.1569
```

Nyt testin $p$-arvo on $p = 0.47$, joka on se todennäköisyys, että nollahypoteesin mukaisesta populaatiosta saadaan $z$-tunnusluvun arvo, joka on itseisarvoltaan vähintään yhtä suuri kuin aineistosta laskettu $z$-tunnusluvun itseisarvo. Testin voima todelliselle parametrinarvolle on (taas) noin 0.16. Ennen aineiston simulointia todennäköisyys, että testi tulee hylkäämään nollahypoteesin $\mu = \frac{1}{2}$ oli 0.16. Tässä tapauksessa testin voima on niin pieni, että hyväksymispäätöstä ei pitäisi tulkita todisteena $H_0$:n puolesta, mikäli $\mu = 0.2012$ on käytännön eli planeetan Z ilmaston kannalta merkittävästi erilainen kuin arvo $\mu_0 = \frac{1}{2}$.

Tämän tarinan eräänä opetuksena voidaan pitää sitä, että pienellä otoskoolla testin voima on helposti niin pieni, että emme voi sen avulla havaita käytännön kannalta tärkeitä eroja nollahypoteesista. Tällaisessa tilanteessa pitäisi yrittää hankkia suurempi otos.

## 6.6 Testien ja luottamusvälien duaalisuus

Tarkastelemme esimerkin vuoksi kaksisuuntaisen $z$-testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

hyväksymisaluetta, eli niitä $\mu_0$ joita tason $\alpha$ testi (6.8) ei hylkää. Testi ei hylkää, mikäli $|z| \leq z_{\alpha/2}$ eli mikäli

$$\begin{aligned} & -z_{\alpha/2} \leq \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\ \Leftrightarrow \quad & \bar{y} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu_0 \leq \bar{y} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \end{aligned}$$

**Taulukko 6.1** Merkitsevyystason $0 < \alpha < 1$ testejä ja luottamusvälejä normaalijakautuneelle populaatiolle $N(\mu, \sigma^2)$, kun varianssi $\sigma^2$ on tunnettu. Tässä $z = (\bar{y} - \mu_0)/(\sigma/\sqrt{n})$, ja $Z \sim N(0,1)$ ja $z_u$ on $N(0,1)$-jakauman $u$-yläkvantiili.

| $H_0$ | $H_1$ | Hylkäysalue | $p$-arvo | Luottamusväli |
|---|---|---|---|---|
| $\mu \leq \mu_0$ | $\mu > \mu_0$ | $z > z_\alpha$ | $P(Z \geq z)$ | $[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$ |
| $\mu \geq \mu_0$ | $\mu < \mu_0$ | $z < -z_\alpha$ | $P(Z \leq z)$ | $(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}]$ |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert z\rvert > z_{\alpha/2}$ | $P(\lvert Z\rvert \geq \lvert z\rvert)$ | $[\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$ |

mutta tunnistamme, että alempi epäyhtälöpari määrittelee tason $1 - \alpha$ kaksisuuntaisen $z$-luottamusvälin. Toisin sanoen, $z$-testi ei hylkää nollahypoteesia $\mu = \mu_0$ täsmälleen silloin, kun $\mu_0$ kuuluu luottamustason $1 - \alpha$ luottamusväliin (5.10).

Voimme sanoa, että kaksisuuntainen $z$-luottamusväli saadaan kääntämällä kaksisuuntaisen $z$-testin hyväksymisalue. Tarkemmin sanoen, ratkaisemme kaikkien muotoa $H_0 : \mu = \mu_0$ olevien kaksisuuntaisten testien hyväksymisalueet. Voimme samaan tapaan kääntää myös yksisuuntaisten $z$-testien hyväksymisalueet, ja näin menetellen saadaan taulukossa 6.1 luetellut tapaukset.

Tässä valossa testin $p$-arvolle voidaan antaa uusi tulkinta. Mikäli $p > \alpha$, niin tällöin $\mu_0$ kuuluu testiä vastaavaan luottamustason $1 - \alpha$ luottamusväliin. Jos taas $p < \alpha$, niin $\mu_0$ ei kuulu ko. luottamusväliin.

Merkitsevyystason $\alpha$ testit ja luottamustason $1 - \alpha$ luottamusvälit yrittävät antaa vastauksen samantapaiseen kysymykseen, mutta erilaisista näkökulmista. Testissä kiinnitetään yksi parametrinarvo, ja tutkitaan ovatko havainnot sopusoinnussa tämän parametrinarvon kanssa. Luottamusväli yrittää kertoa suoraan, mitkä parametrinarvot ovat sopusoinnussa havaintojen kanssa. Tähän testien ja luottamusvälien vastaavuuteen voidaan viitata sanomalla, että ne ovat duaalisia käsitteitä.

Planeetan Z esimerkissä yksisuuntaista hypoteesia

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

vastaa parametrin $\mu$ yksisuuntainen 95 %:n luottamusväli $[0.20, \infty)$, joka ei sisällä esimerkissä kiinnostavaa arvoa 0, joten nollahypoteesi $H_0 : \mu \leq 0$ hylätään merkitsevyystasolla 0.05. Testin $p$-arvo on $p = 0.011 < 0.05$, joten myös tästä nähdään, että $\mu_0 = 0$ ei kuuluu ko. 95 %:n yksisuuntaiseen luottamusväliin ilman että luottamusväliä edes tarvitsee laskea.

Kaksisuuntaista hypoteesia

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

vastaa parametrin $\mu$ kaksisuuntainen 95 %:n luottamusväli $[0.10, 1.35]$, joka sisältää esimerkissä kiinnostavan arvon $\frac{1}{2}$, joten nollahypoteesi $\mu = \frac{1}{2}$ hyväksytään merkitsevyystasolla 0.05. Testin $p$-arvo on $p = 0.47$, mistä myöskin nähdään, että $\mu_0 = \frac{1}{2}$ kuuluu kaksisuuntaiseen 95 %:n luottamusväliin.

Testaaminen johtaa käyttäjän helposti mustavalkoiseen ajatteluun: nollahypoteesi joko hyväksytään tai hylätään. Tällöin käyttäjän huomio kiinnittyy pois siitä, kuinka epävarmaa aineiston antama informaatio parametrista on.

**Taulukko 6.2** Merkitsevyystason $0 < \alpha < 1$ testejä ja luottamusvälejä normaalijakautuneen populaation $N(\mu, \sigma^2)$ odotusarvolle $\mu$, kun myös varianssi $\sigma^2$ on tuntematon. Tässä $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$, $s$ on otoskeskihajonta, $T \sim t_{n-1}$ ja $t_{n-1}(u)$ on $t_{n-1}$-jakauman $u$-yläkvantiili.

| $H_0$ | $H_1$ | Hylkäysalue | $p$-arvo | Luottamusväli |
|---|---|---|---|---|
| $\mu \leq \mu_0$ | $\mu > \mu_0$ | $t > t_{n-1}(\alpha)$ | $P(T \geq t)$ | $[\bar{y} - t_{n-1}(\alpha)\frac{s}{\sqrt{n}}, \infty)$ |
| $\mu \geq \mu_0$ | $\mu < \mu_0$ | $t < -t_{n-1}(\alpha)$ | $P(T \leq t)$ | $(-\infty, \bar{y} + t_{n-1}(\alpha)\frac{s}{\sqrt{n}}]$ |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert t\rvert > t_{n-1}(\frac{\alpha}{2})$ | $P(\lvert T\rvert \geq \lvert t\rvert)$ | $[\bar{y} - t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}]$ |

Sen sijaan luottamusväli kvantifioi epävarmuuden selkeällä tavalla. Kun pisteestimaatti sekä luottamusväli lasketaan käytännön kannalta kiinnostavalle parametrille, niin saadaan tuloksia, jotka voidaan tulkita suoraan. Sen sijaan asiantuntemattomat lukijat tulkitsevat testien tulokset toisinaan aivan nurinkurisella tavalla. Testauksessa tutkijalla pitäisi olla selkeä käsitys testin voimafunktiosta sellaisilla parametrinarvoilla, jotka ovat käytännön kannalta merkityksellisiä.

Vaikka testit ja luottamusvälit ovat duaalisia käsitteitä, niin sellaisissa yksinkertaisissa tilanteissa joissa molemmat lähestymistavat ovat mahdollisia *luottamusvälien laskeminen on parempi tapa analysoida aineistoa kuin testaaminen.*

## 6.7 Normaalijakautuneen populaation odotusarvon testaus, kun myös varianssi on tuntematon

Jos sekä odotusarvo että varianssi ovat tuntemattomia, niin testit perustetaan saranasuureelle

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}},$$

jolla on $t$-jakauma vapausasteluvulla $n - 1$, mikäli $\mu = \mu_0$. Tästä tiedosta saadaan johdettua $t$-testit eri tapauksille matkimalla $z$-testien johtoa. Tulokset on koottu taulukkoon 6.2. Käytännössä $t$-testi suoritetaan aina jollakin tarkoitukseen sopivalla tietokoneohjelmalla. Esim. R-ohjelmistosa kaikki taulukon 6.2 tulokset saadaan laskettua vaivattomasti funktiolla `t.test`.

Jos esimerkiksi vastahypoteesi on kaksisuuntainen

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0,$$

ja $\mu_0$ on kiinteä luku, niin testisuure lasketaan kaavalla

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}},$$

jossa $s$ on otoskeskihajonta. Vastaavalla satunnaismuuttuja $t(\mathbf{Y})$ noudattaa $t$-jakaumaa $n-1$ vapausasteella, ja testisuureen sekä pienet että suuret arvot ovat kriittisiä. Testi hylkää nollahypoteesin, mikäli

$$|t| > t_{n-1}(\frac{\alpha}{2}),$$

ja sen $p$-arvo on

$$p = P(|T| \geq |t|) = 2(1 - F_{n-1}(|t|)),$$

jossa satunnaismuuttuja $T \sim t_{n-1}$, ja $F_\nu$ tarkoittaa $t_\nu$-jakauman kertymäfunktiota. Kaksisuuntainen $t$-luottamusväli saadaan ratkaisemalla kaikki ne $\mu_0$-arvot, joilla $t$-testi hyväksyisi nollahypoteesin $H_0 : \mu = \mu_0$ kaksisuuntaiselle vastahypoteesille:

$$\begin{aligned} & \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \leq t_{n-1}\left(\frac{\alpha}{2}\right) \\ \Leftrightarrow \quad & -t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\mu_0 - \bar{y}}{s/\sqrt{n}} \leq t_{n-1}\left(\frac{\alpha}{2}\right) \\ \Leftrightarrow \quad & \bar{y} - \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \leq \mu_0 \leq \bar{y} + \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \end{aligned}$$

Muut taulukon 6.2 rivit saadaan järkeiltyä samaan tapaan.

Voimafunktion laskeminen $t$-testille on monimutkaisempaa kuin $z$-testille, mutta tämä kuitenkin onnistuu ns. epäkeskisen $t$-jakauman avulla. Tuloksena saadaan samantapaisia käyriä kuin $z$-testin voimafunktiolle.

## 6.8 Binomijakauman parametrin testaus

Jos tahdotaan testata lantin harhattomuutta, niin testi voidaan perustaa suoraan onnistumisten lukumäärälle $X \sim \text{Bin}(n, p)$. (Onnistuminen voi nyt olla yhtä kuin kruunan saaminen yhdellä lantin heitolla.) Tässä tapauksessa hypoteesit ovat

$$H_0 : p = \frac{1}{2}, \qquad H_1 : p \neq \frac{1}{2}.$$

Testi voidaan perustaa sille idealle, että tutkitaan havainnon poikkeamaa nollahypoteesin mukaisen jakauman odotusarvosta $EX = n/2$. Kaksisuuntaisen testin $p$-arvo voidaan laskea kaavalla

$$p = P_{1/2}\left(\left|X - \frac{n}{2}\right| \geq \left|x - \frac{n}{2}\right|\right).$$

Tässä alaindeksi $1/2$ tarkoittaa sitä, että oletamme nollahypoteesin mukaisesti, että $X \sim \text{Bin}(n, 1/2)$, ja $x$ on havaittu onnistumisten lukumäärä.

Tämän testin $p$-arvo saadaan laskettua karkeasti käyttämällä normaaliapproksimaatiota, jonka mukaan nollahypoteesin vallitessa satunnaismuuttujalla

$$\frac{X - EX}{\sqrt{\text{var}\, X}} = \frac{X - \frac{n}{2}}{\sqrt{n\,\frac{1}{2}\,\frac{1}{2}}}$$

on osapuilleen standardinormaalijakauma $N(0, 1)$. Tämän takia $p$-arvo saadaan osapuilleen kaavalla

$$2\left(1 - \Phi\left(\frac{2}{\sqrt{n}}\left|x - \frac{n}{2}\right|\right)\right),$$

tai kaavalla

$$2\left(1 - \Phi\left(\frac{2}{\sqrt{n}}\left(\left|x - \frac{n}{2}\right| - \frac{1}{2}\right)\right)\right),$$

Jälkimmäisessä kaavassa tehtiin jatkuvuuskorjaus, ja $\Phi$ on $N(0,1)$-jakauman kertymäfunktio.

Jos $n = 1000$ ja onnistumisia on tullut $x = 460$, niin nämä $p$-arvon normaaliapproksimaatiot ovat seuraavat.

```
n <- 1000
x <- 460
2 * pnorm(lower = FALSE, 2/sqrt(n) * abs(x - n/2))

## [1] 0.01141

2 * pnorm(lower = FALSE, 2/sqrt(n) * (abs(x - n/2) - 1/2))

## [1] 0.01248
```

Binomijakauman häntätodennäköisyydet saadaan laskettua tarkasti tietokoneohjelmilla. Esimerkiksi, jos $n = 1000$ ja onnistumisia on tullut $x = 460$, niin testin $p$-arvo saadaan selvitettyä ilman approksimaatioita komennolla

```
binom.test(460, 1000, p = 0.5)

##
##  Exact binomial test
##
## data:  460 and 1000
## number of successes = 460, number of trials = 1000, p-value =
## 0.01244
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4288 0.4915
## sample estimates:
## probability of success
##                   0.46
```

Voimme samaan tapaan käsitellä muutkin kaksisuuntaiset testit

$$H_0 : p = p_0, \qquad H_1 : p \neq p_0,$$

ja kääntämällä näiden testien hyväksymisalueet saadaan jaksossa 5.8 mainittu Clopperin ja Pearsonin tarkka luottamusväli parametrille $p$.

Yksisuuntaiset testit saadaan käsiteltyä samaan tapaan.

## 6.9 $p$-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa

$P$-arvoa voidaan ajatella mittana sille, miten hyvin havainto on sopusoinnussa nollahypoteesin kanssa. Koska tämä käsite on vaikeatajuinen, tilastotieteen soveltajilla on kaikenlaisia harhakäsityksiä siitä, mitä $p$-arvo tarkoittaa.

Eräs yleinen harhakäsitys on se, että $p$-arvo on todennäköisyys sille, että nollahypoteesi pitää paikkansa. Jotta tällaiseen täysin virheelliseen käsitykseen voisi päätyä, pitää tehdä monta räikeää käsitteellistä virhettä, kuten esimerkiksi seuraavat.

1. Unohdetaan, että toimitaan frekventistisen tilastotieteen puitteissa. Frekventistisessä tilastotieteessä parametrinarvoihin tai tilastollisiin hypoteeseihin ei liitetä todennäköisyyksiä.
2. Tämän sekaannuksen lisäksi ajatellaan, että $P(H_0 \mid Y = y)$ on sama asia kuin $P(Y = y \mid H_0)$.
3. Lopuksi vielä ajatellaan, että $p$-arvo on sama asia kuin $P(Y = y \mid H_0)$, mitä se ei ole. $P$-arvo on häntätodennäköisyys, tarkemmin sanoen todennäköisyys sille, että nollahypoteesin mukaisesta populaatiosta saadaan arvo, joka on yhtä kummallinen tai vielä kummallisempi kuin havaittu arvo.

Testin $p$-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa, vaikka näin lukijalle kerrotaan lukuisissa tilastotieteen soveltajille tarkoitetussa oppikirjoissa.

## 6.10 Tilastollisten testien väärinkäyttöä

Monille tilastotieteen soveltajille on syntynyt sellainen mielikuva, että kokeellisen tutkimuksen päämäränä on laskea $p$-arvo jollekin testille. Jos $p$-arvo on riittävän pieni (esim. $p < 0.05$), niin tuloksen saa julkaistua jossakin alan lehdessä. Jos $p$-arvo ei ole riittävän pieni, tutkimusta ei kannata lähettää arvioitavaksi, koska sitä ei kuitenkaan tulla julkaisemaan. Valitettavasti tämä harha ei koske yksinomaan yksittäisiä tutkijoita, vaan tällainen käsitys on ollut yleinen myös vaikutusvaltaisten lehtien arvioijien ja toimittajien parissa. Tällainen käytäntö johtaa *julkaisuharhaan* (engl. *publication bias*): kirjallisuudessa julkaistaan enimmäkseen nollahypoteesin hylkääviä tutkimuksia riippumatta siitä, mikä todellisuudessa on asian laita. Tällainen käytäntö perustuu väärinkäsityksiin, rituaaleihin ja taikauskoon eikä sillä ole mitään tekemistä kunnollisen tieteellisen tutkimuksen kanssa eikä kunnollisen tilastotieteen soveltamisen kanssa (ks. esim. [3] tai [2]).

Lisäksi useimmiten julkaisuissa testataan nollahypoteeseja, joista jo ennen tutkimuksen tekoa tiedetään, että ne eivät voi pitää paikkaansa. Nämä ovat ns. hölmöjä nollahypoteeseja (engl. *silly null*). Käsittelyllä on todellisuudessa kuitenkin aina jokin vaikutus, joten nollahypoteesi $\mu = 0$ (ei vaikutusta) ei voi pitää kirjaimellisesti paikkaansa, eikä kukaan oikeasti usko tätä tarkkaa, pistemäistä (engl. *sharp null*, *point null*) nollahypoteesia. Jos paikkansa pitämätöntä pistemäistä nollahypoteesia ei saada testillä hylättyä, niin syynä on se, että testillä ei ollut riittävästi voimaa, eli otoskoko oli liian pieni.

Jos taas otoskokoa kasvatetaan, ja vihdoin pystytään nollahypoteesi hylkäämään, niin voi olla, että aineiston nojalla arvioitu vaikutuksen suuruus on niin pieni, että sillä ei ole mitään käytännön merkitystä.

> *Kysymys:* Jos nollahypoteesin mukaan $\mu = 0$, ja paras estimaattimme vaikutuksen suuruudelle on $\hat{\mu} = 0.1$, niin onko tällä erolla käytännössä merkitystä?

Tämä on kysymys, johon tilastotiede ei pysty antamaan vastausta. Vastauksen pitää tulla substanssialan asiantuntijalta. (Mitä asiaa mitattiin? Mitä yksikköjä käytettiin? jne.) Tilastollinen merkitsevyys (engl. *statistical significance*) ja käytännön merkittävyys (engl. *practical significance*) ovat aivan eri asioita.

On hedelmällisempää ja informatiivisempaa yrittää estimoida vaikutuksen suuruutta ja yrittää kvantifioida estimaattiin liittyvää epävarmuutta (keskivirhe, luottamusväli!) kuin yrittää testata, onko vaikutus nolla.

Tilastotieteen soveltajien intoon testata kaikkea mahdollista viitataan usein lyhenteellä NHST (*null hypothesis significance testing*). Hakusanan NHST avulla on helppo löytää tämän käytännön kritiikkiä.

Joillakin tilastotieteeseen vahvasti nojaavilla aloilla (esim. lääketiede, terveystieteet ja psykologia) on käynnissä uudistusliike, jossa pyritään pois epätarkoituksenmukaisesta hypoteesien testauksesta. Tämän sijasta

- lasketaan piste-estimaatteja ja väliestimaatteja vaikutuksen suuruudelle,
- yhdistetään aikaisempien tutkimusten tuloksia, eli harrastetaan meta-analyysia.

Jotkut kirjoittajat käyttävät tästä uudistusliikkeestä nimitystä uusi tilastotiede (engl. *new statistics*) [1] — tilastotieteen näkökulmasta uudessa tilastotieteessä ei ole paljoa uutta.

Kerrataan lopuksi vielä seuraavat asiat:

- Nollahypoteesin hyväksyminen testissä ei tarkoita sitä, että oltaisiin löydetty todisteita nollahypoteesin puolesta. Se tarkoittaa sitä, että ei olla löydetty riittävän painavia todisteita nollahypoteesia vastaan.
- Testin tekemän päätöksen voi lukea $p$-arvosta, joka mittaa sitä kuinka hyvin aineisto on sopusoinnussa nollahypoteesin kanssa.
- Testin $p$-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa.
- Testaamisen sijasta kannattaa laskea piste-estimaatteja ja luottamusvälejä, mikäli tämä on mahdollista.

# Kirjallisuutta

[1] Geoff Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, 2012.

[2] Esa Läärä. Statistics: reasoning on uncertainty, and the insignificance of test null. *Annales Zoologici Fennica*, 46:138–157, 2009.

[3] John A. Nelder. Statistics for the millennium: From statistics to statistical science. *The Statistician*, 48:257–269, 1999.

# Luku 7

# Kahden populaation vertaaminen

Tässä luvussa tarkastelemme frekventistisen tilastotieteen keinoin tilanteita, joissa vertaillaan kahden populaation odotusarvoparametrien suuruuksia populaatiosta saatujen otosten perusteella. Rajoitumme tapaukseen, jossa populaatiot oletetaan normaalijakautuneiksi.

## 7.1 Kahden populaation vertailu, kun otosten välillä on yhteyttä

Usein koeasetelma tuottaa mittauspareja $(y_{1i}, y_{2i})$, jonka komponentit ovat keskenään samankaltaisia. Esimerkiksi, jos mittaukset $y_{1i}$ ja $y_{2i}$ tehdään kaikilla $i$ samasta otosyksiköstä (esim. samasta henkilöstä) ennen ja jälkeen käsittelyn, niin silloin vastaavia satunnaismuuttujia $Y_{1i}$ ja $Y_{2i}$ ei voida pitää riippumattomina, vaan samaan otosyksikköön $i$ liittyvät muuttujat $Y_{1i}$ ja $Y_{2i}$ ovat samankaltaisempia kuin eri yksikköihin $i$ ja $j$ liittyvät muuttujat $Y_{1i}$ ja $Y_{2j}$. Samanlainen tilanne syntyy, jos $y_{1i}$ ja $y_{2i}$ saadaan eri otosyksiköistä, jotka on kuitenkin valittu sillä tavalla, että ne ovat jonkin attribuutin mukaan samankaltaisia. Tarkastelemme nyt tällaisten toisistaan riippuvien otosten eli *parittaisten* (engl. *paired, related, matched*) otosten analysiointia.

Tällainen tilanne voidaan käsitellä tarkastelemalla erotuksia

$$d_i = y_{1i} - y_{2i}, \qquad i = 1, \ldots, n$$

Mikäli vastaavia satunnaismuuttujia

$$D_i = Y_{1i} - Y_{2i}, \qquad i = 1, \ldots, n$$

voidaan pitää riippumattomina, samoinjakautuneina ja (ainakin likimäärin) normaalijakautuneina,

$$D_i \sim N(\delta, \sigma^2), \qquad i = 1, \ldots, n,$$

niin tällöin populaatioiden odotusarvojen erotus on

$$\delta = \mu_1 - \mu_2,$$

ja tyypillisesti $\sigma^2$ on tuntematon. Odotusarvoparametrien erotusta $\delta = \mu_1 - \mu_2$ voidaan nyt analysoida soveltamalla $t$-luottamusväliä tai $t$-testiä erotuksiin $d_i$.

Tässä ns. *parittaisessa t-luottamusvälissä* tai *parittaisessa t-testissä* ei tarvitse esim. olettaa, että populaatioilla $Y_{1i}$ ja $Y_{2i}$ olisi sama varianssi (kuten seuraavassa jaksossa tehdään), vaan jakaumaoletukset tehdään erotuksille $D_i$. Jakaumaoletus on voimassa esim. silloin, kun kaksikomponenttiset vektorit $(Y_{1i}, Y_{2i})$ ovat satunnaisotos jostakin kaksiulotteiseta normaalijakaumasta.

## 7.2 Kaksi riippumatonta otosta normaalijakautuneista populaatioista

Tarkastelemme tilannetta, joka mallinnetaan kahdella riippumattomalla satunnaisotoksella normaalijakaumista $N(\mu_1, \sigma_1^2)$ ja $N(\mu_2, \sigma_2^2)$. Populaatiosta 1 saadaan $n_1$ havaintoa $y_{1i}$ ja populaatiosta 2 saadaan $n_2$ havaintoa $y_{2i}$. Tavoitteena on verrata populaatioiden odotusarvoja $\mu_1$ ja $\mu_2$, jotka ovat tuntemattomia parametreja. Kehitämme tätä varten sekä luottamusvälejä että testejä.

Tällä tavalla voitaisiin mallintaa koetilanne, jossa tehdään mittauksia populaatiosta 1, jonka yksilöihin kohdistetaan käsittely 1 sekä populaatiosta 2, jonka yksilöihin kohdistetaan käsittely 2, mikäli kaikki yksilöt ovat toisistaan erillisiä. Mikäli tämä on käytännössä mahdollista, populaatiot mielellään muodostetaan satunnaistamalla, eli jakamalla havaintoyksiköt satunnaisesti kahteen ryhmään.

Oletus kahdesta riippumattomasta satunnaisotoksesta tarkoittaa sitä, että oletamme havaintoja vastaavien satunnaismuuttujien $Y_{ki}$ noudattavan jakaumia

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \sim N(\mu_1, \sigma_1^2) \tag{7.1}$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \sim N(\mu_2, \sigma_2^2), \tag{7.2}$$

ja että kaikki satunnaismuuttujat $Y_{ki}$ ovat riippumattomia. Otoskoot $n_1$ ja $n_2$ voivat olla erisuuria.

Populaatioiden parametreja $(\mu_1, \sigma_1^2)$ ja $(\mu_2, \sigma_2^2)$ voidaan estimoida tuttuun tapaan otoskeskiarvolla ja otosvarianssilla siten, että populaation $k$ parametrit estimoidaan populaatiosta $k$ saadusta otoksesta. Osoitamme alaindeksillä, kummasta populaatiosta otossuureet on laskettu. Käytämme merkintöjä

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}, \qquad \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2, \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$$

Merkitsemme näitä estimaatteja vastaavia satunnaismuuttujia (ts. estimaattoreita) estimaatteja vastaavilla suurilla kirjaimilla

$$\bar{Y}_1, \quad \bar{Y}_2, \quad S_1^2 \quad \text{ja} \quad S_2^2.$$

Tiedämme jakson 4.4.2 kaavojen (4.14)–(4.16) perusteella, että

$$\bar{Y}_1 \sim N(\mu_1, \frac{1}{n_1}\sigma_1^2) \qquad \bar{Y}_2 \sim N(\mu_2, \frac{1}{n_2}\sigma_2^2) \tag{7.3}$$

$$\frac{n_1 - 1}{\sigma_1^2}S_1^2 \sim \chi^2_{n_1 - 1} \qquad \frac{n_2 - 1}{\sigma_2^2}S_2^2 \sim \chi^2_{n_2 - 1}, \tag{7.4}$$

ja että lisäksi toisaalta $\bar{Y}_1$ ja $S_1^2$ ovat riippumattomia ja että $\bar{Y}_2$ ja $S_2^2$ ovat riippumattomia satunnaismuuttujia. Nyt itseasiassa kaikki neljä satunnaismuuttujaa ovat riippumattomia sillä perusteella, että riippumattomista otoksista johdettavat estimaattorit ovat keskenään riippumattomia.

Kinnostuksen kohteena on populaatioiden odotusarvojen erotus

$$\delta = \mu_1 - \mu_2,$$

ja sitä estimoidaan vastaavalla otoskeskiarvojen erotuksella,

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2. \tag{7.5}$$

Vastaavalla estimaattorilla on normaalijakauma sen takia, että riippumattomien normaalijakaumaa noudattavien satunnaismuuttujien lineaarikombinaatio tunnetusti noudattaa aina normaalijakaumaa. Laskemalla erotuksen odotusarvo ja varianssi saadaan johdettua kyseisen normaalijakauman parametrit, ja tällä tavalla nähdään, että

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2) \tag{7.6}$$

Tässä vaiheessa joudutaan erilaisiin tarkasteluihin sen mukaan, mitä populaatioiden variansseista oletetaan.

## Varianssit tunnettuja

Jos molemmat varianssiparametrit $\sigma_1^2$ ja $\sigma_2^2$ ovat tunnettuja vakioita, niin satunnaismuuttuja

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{\sqrt{\frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2}}$$

noudattaa standardinormaalijakaumaa $N(0,1)$. Tällä tavalla saadaan tuttuun tapaan johdettua luottamusväli odotusarvojen erotukselle $\delta = \mu_1 - \mu_2$ tai voidaan johtaa testit yksisuuntaiselle hypoteeseille

$$H_0 : \delta \leq \delta_0, \quad H_1 : \delta > \delta_0$$

tai yksisuuntaiselle hypoteesille

$$H_0 : \delta \geq \delta_0, \quad H_1 : \delta < \delta_0$$

tai kaksisuuntaiselle hypoteesille

$$H_0 : \delta = \delta_0, \qquad H_1 : \delta \neq \delta_0.$$

Huomaa, että tässä tarkka hypoteesi $H_0 : \delta = \delta_0$ on oikeasti yhdistetty hypoteesi, koska se vastaa parametriavaruuden osajoukkoa

$$\Theta_0 = \{(\mu_1, \mu_2) : \mu_1 - \mu_2 = \delta_0\}.$$

## Varianssit yhtäsuuria, mutta tuntemattomia

Edellistä käyttökelpoisempi tilanne on se, jossa populaatioiden varianssit ovat tuntemattomia, mutta ne oletetaan yhtäsuuriksi. Ts. oletetaan, että

$$\sigma_1^2 = \sigma_2^2 = \sigma^2,$$

jossa $\sigma^2$ on tuntematon parametri. Tällöin mallin parametrivektori on

$$(\mu_1, \mu_2, \sigma^2).$$

Kiinnostuksen kohteena on odotusarvojen erotus $\delta = \mu_1 - \mu_2$. Tämä tilanne on erikoistapaus ns. *varianssianalyysistä* (engl. *analysis of variance, ANOVA*), johon voi tutustua tarkemmin lineaaristen mallien kursseilla tai oppikirjoista.

Avainajatus analyysissä on muodostaa yhteiselle varianssille $\sigma^2$ sellainen estimaattori $S_p^2$, jonka jakauman hallitsemme, ja joka käyttää hyväksi molempien otoksien sisältämän informaation varianssista $\sigma^2$. Tämän jälkeen osaamme laskea parametrin $\delta$ estimaatin keskivirheen, ja loppu on tuttujen ideoiden soveltamista.

Käytämme hyväksi $\chi^2$-jakauman ominaisuuksia. Jos $X \sim \chi^2_\nu$, niin sen odotusarvo on

$$EX = \nu, \tag{7.7}$$

mikä voidaan päätellä esim. gammajakauman odotusarvon kaavan avulla, sillä $\chi^2_\nu$ jakauma on gammajakauma Gamma($\frac{1}{2}\nu, \frac{1}{2}$). Tarvitsemme myös $\chi^2$-jakauman yhteenlaskuominaisuutta. Jos

$$X_1 \sim \chi^2_{\nu_1}, \quad X_2 \sim \chi^2_{\nu_2}, \quad X_1 \perp\!\!\!\perp X_2,$$

niin tällöin

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2} \tag{7.8}$$

Tämä voidaan johtaa esim. gammajakauman yhteenlaskuominaisuudesta.

Tietojen (7.3) sekä khiin neliön jakauman yhteenlaskuominaisuuden nojalla

$$\frac{n_1 - 1}{\sigma^2} S_1^2 + \frac{n_2 - 1}{\sigma^2} S_2^2 \sim \chi^2_{n_1+n_2-2},$$

ts. on voimassa

$$\frac{n_1 + n_2 - 2}{\sigma^2} S_P^2 \sim \chi^2_{n_1+n_2-2}, \tag{7.9}$$

kun määrittelemme *yhdistetyn* varianssiestimaattorin $S_p^2$ (engl. *pooled variance estimator*) seuraavana estimaattorien $S_1^2$ ja $S_2^2$ lineaarikombinaationa,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{7.10}$$

$$= \frac{1}{n_1 + n_2 - 2} \left[ \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \right] \tag{7.11}$$

Yhdistetty varianssiestimaattori on harhaton, sillä jakaumatulosta (7.9) sekä $\chi^2$-jakauman odotusarvon kaavaa käyttämällä

$$ES_p^2 = E\left[ \frac{\sigma^2}{n_1 + n_2 - 2} \frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 \right] = \frac{\sigma^2}{n_1 + n_2 - 2} (n_1 + n_2 - 2) = \sigma^2.$$

Kun otetaan tuloksen (7.9) lisäksi huomioon se seikka, että

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2\right)$$

niin nähdään, että seuraavalla satunnaismuuttujalla on $t$-jakauma vapausasteluvulla $n_1 + n_2 - 2$,

$$t(\mathbf{Y}) = \frac{\left(\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)\right) / \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,\sigma\right)}{S_p/\sigma} \sim t_{n_1+n_2-2}.$$

Sieventämällä nähdään, että

$$t(\mathbf{Y}) = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \tag{7.12}$$

joten $t(\mathbf{Y})$ on saranasuure kiinnostavalle parametrille $\delta = \mu_1 - \mu_2$. Luottamusvälit ja testit voidaan perustaa tälle tulokselle.

Luottamustason $0 < 1 - \alpha < 1$ kaksisuuntainen luottamusväli saadaan johdettua tuttuun tapaan lähtemällä liikkeelle tuloksesta

$$P_{(\mu_1,\mu_2,\sigma^2)}\left(\left|\frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \leq t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Ratkaisemalla tämä epäyhtälö tuntemattoman $\delta$ suhteen nähdään, että jokaisessa parametriavaruuden pisteessä pätee todennäköisydellä $1 - \alpha$ paikkansa kaksoisepäyhtälö

$$\bar{Y}_1 - \bar{Y}_2 - t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \delta \leq$$
$$\bar{Y}_1 - \bar{Y}_2 + t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ts. luottamustason $1 - \alpha$ kaksisuuntainen luottamusväli saadaan laskemalla aineistosta parametrin $\delta = \mu_1 - \mu_2$ estimaatti

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2 \tag{7.13}$$

sekä yhdistetty varianssiestimaatti

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{7.14}$$

minkä jälkeen luottamusväli on muotoa

$$\text{estimaatti} \pm (t\text{-jakauman kriittinen piste}) \times \text{estimaatin keskivirhe}$$

eli tarkemmin sanoen se on

$$\left[\hat{\delta} - t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \hat{\delta} + t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{7.15}$$

Merkitsevyystason $0 < \alpha < 1$ kaksisuuntainen testi tarkalle hypoteesille

$$H_0 : \delta = \delta_0, \qquad H_1 : \delta \neq \delta_0$$

saadaan suoritettua laskemalla aineistosta testisuure

$$t = t(\mathbf{y}) = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{7.16}$$

jota verrataan $t$-jakaumaan vapausasteluvulla $n_1 + n_2 - 2$. Testi hylkää nollahypoteesin täsmälleen silloin, kun

$$|t| > t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right).$$

Yksisuuntainen testi hypoteeseille

$$H_0 : \delta \leq \delta_0, \qquad H_1 : \delta > \delta_0$$

hylkää nollahypoteesin, jos

$$t > t_{n_1+n_2-2}(\alpha),$$

ja hypoteeseille

$$H_0 : \delta \geq \delta_0, \qquad H_1 : \delta < \delta_0$$

hylkää nollahypoteesin, jos

$$t < -t_{n_1+n_2-2}(\alpha),$$

Tavallisesti näissä testeissä $\delta_0 = 0$. Usein testataan tarkkaa nollahypoteesia $\mu_1 - \mu_2 = 0$, jonka mukaan populaatiolla on sama odotusarvo. Huomaa, että tämä tarkka hypoteesi on yhdistetty, sillä se vastaa parametriavaruuden osajoukkoa

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) : \mu_1 = \mu_2, \sigma^2 > 0\}$$

## Varianssit erisuuria ja tuntemattomia

Jos varianssit ovat tuntemattomia, ja ne eivät ole yhtäsuuria, niin ratkaistavana on ns. Behrensin–Fisherin ongelma, jolle ei löydy tarkkaa ratkaisua. Sen sijaan on löydetty likimääräisiä ratkaisuja.

Esim. R:n funktio `t.test` käyttää tässä tilanteessa ns. Welchin testiä, joka taas perustuu ns. Satterthwaiten approksimaatioon. R käyttää kahden populaation vertailuun Welchin testiä, ellei sitä pyydetä erikseen olettamaan, että varianssit ovat yhtäsuuret.

# Luku 8

# Yhteensopivuuden ja riippumattomuuden testaaminen

Tässä kappaleessa tarkastelemme eräitä kuuluisia frekventistisen tilastotieteen testejä, joilla voidaan tutkia diskreettien havaintojen sopivuutta erilaisten tilastollisten mallien kanssa. Kaikki tämän kappaleen testit ovat likimääräisiä, ja niiden käyttö vaatii suurta otoskokoa.

## 8.1 Pearsonin testisuure

Tilastollinen malli koostuu diskreeteistä satunnaismuuttujia $Y_1, \ldots, Y_n$, joista kukin voi saada minkä tahansa arvoista $1, 2, \ldots, k$. Oletamme, että satunnaismuuttujat $Y_h$ ovat riippumattomia ja samoin jakautuneita, jolloin niiden yhteisjakauma tiedetään, mikäli tunnetaan eri vaihtoehtojen $1, \ldots, k$ eli eri luokkien todennäköisyydet

$$p_i = P(Y_h = i), \qquad i = 1, \ldots, k \tag{8.1}$$

Luokkien todennäköisyyksien summa on yksi, joten mallissa on $k - 1$ vapaata parametria, joiksi voidaan valita $k - 1$ ensimmäisen luokan todennäköisyydet $p_1, \ldots, p_{k-1}$. Tämän jälkeen $p_k$ voidaan laskea muiden $p_i$ funktiona kaavalla

$$p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}. \tag{8.2}$$

Binomikoe on tämän mallin erikoistapaus, jossa vaihtoehtoja on vain kaksi, ja joista yhtä pidetään onnistumisena ja toista epäonnistumisena. Tämän jakson mallia voidaan kutsua multinomikokeeksi.

Olkoon $n_i$ luokan $i$ havaittu frekvenssi, eli $n_i$ on niiden indeksien $h$ lukumäärä, joille $y_h = i$. Merkitsemme vastaavia satunnaismuuttujia symboleilla $N_i$. Binomijakauman määritelmän perusteella

$$N_i \sim \mathrm{Bin}(n, p_i), \qquad i = 1, \ldots, k,$$

joten $EN_i = np_i$.

Jos $i$ ja $j$ ovat eri luokkia, niin $N_i$ ja $N_j$ ovat riippuvia satunnaismuuttujia. Esim. jos binomikokeessa onnistumista kutsutaan luokaksi yksi ja epäonnistumista luokaksi kaksi, niin onnistumisten lukumäärän $N_1$ jakauma ja epäonnistumisten lukumäärän $N_2$ jakauma on

$$N_1 \sim \mathrm{Bin}(n, p), \qquad N_2 \sim \mathrm{Bin}(n, 1-p),$$

mutta koska $N_2 = n - N_1$, niin satunnaismuuttujan $N_2$ arvo tiedetään täysin, jos satunnaismuuttujan $N_1$ arvo tiedetään. Jos luokkia on enemmän kuin kaksi, niin frekvenssien keskinäinen riippuvuus ei ole enää yhtä äärimmäistä, mutta riippuvuus säilyy kuitenkin myös tässä tapauksessa. Sen sijaan alla olevat yksittäisten kokeiden lopputulokset $Y_h$ toki ovat riippumattomia. Kun multinomikokeen tuloksia analysoidaan, niin analyysissä pitää ottaa huomioon frekvenssien keskinäinen riippuvuus.

Pearsonin testisuure vertaa luokkien havaittuja frekvenssejä $n_i$ niiden odotettuihin frekvensseihin $np_i$ seuraavalla tavalla:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}. \tag{8.3}$$

Tavanomaisempi tapa mitata vektorin $(n_1, \ldots, n_k)$ ja vektorin $(np_1, \ldots, np_k)$ välistä etäisyyttä olisi esim. laskea vektoreiden euklidinen etäisyys, jolloin ensin summataan komponenttien neliöidyt erotukset $(n_i - np_i)^2$, ja lopuksi tuloksesta lasketaan neliöjuuri. Tällainen vertailu ei tässä tapauksessa ole hyvä ajatus, sillä todennäköisemmissä luokissa on odotettavissa enemmän satunnaisvaihtelua kuin harvinaisemmissa luokissa. Tätä varten Personin testisuureessa erotuksen neliö jaetaan luokan odotetulla frekvenssillä.

Pearsonin testisuure voidaan ilmaista myös kaavalla

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{8.4}$$

jossa $O_i = n_i$ on havaittu frekvenssi (engl. *observed frequency*), ja $E_i = np_i$ on odotettu frekvenssi (engl. *expected frequency*).

Karl Pearson esitti v. 1900 perustelun sille, miksi hänen nimeään kantavalla testisuureella on suuressa otoksessa likimain $\chi^2$-jakauma. Pearsonin esittämä testi on ensimmäinen tunnettu tilastollinen testi, ja se avasi uuden aikakauden tilastollisen päättelyn historiassa.

Yksinkertaisimmassa yhteensopivuustestissä (engl. *goodness of fit test*) testataan yksinkertaista nollahypoteesia

$$H_0 : p_1 = \pi_1, \ldots, p_{k-1} = \pi_{k-1}, p_k = \pi_k, \tag{8.5}$$

jossa luvut $(\pi_i)$ ovat jonkin teorian mukaisia luokkien todennäköisyyksiä, jotka oletetaan tunnetuiksi. Aineistosta lasketaan havaitut frekvenssit $O_i = n_i$. Nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_i = n\,\pi_i, \qquad i = 1, \ldots, k.$$

Suuret Pearsonin testisuureen $X^2$ arvot ovat nollahypoteesille kriittisiä. Nollahypoteesi hylätään merkitsevyystason $0 < \alpha < 1$ testissä, mikäli lasketun testisuureen arvo on suurempi kuin $\chi^2_{k-1}$ jakauman $\alpha$-yläkvantiili. Huomaa, että

$\chi^2_\nu$-jakauman vapausasteluku $\nu$ on tässä täysin määrätyn nollahypoteesin tapauksessa

$$\nu = k - 1, \tag{8.6}$$

eli se on yhtä kuin mallin vapaiden parametrien lukumäärä $k-1$. Testin $p$-arvo on

$$1 - F(t),$$

jossa $F$ on $\chi^2_{k-1}$-jakauman kertymäfunktio, ja $t$ on testisuureen laskettu arvo.

**Esimerkki 8.1** (Nopan harhattomuuden testaaminen) Simuloidaan ensin $n = 2000$ nopanheittoa harhaisesta nopasta, jolle silmälukujen todennäköisyydet ovat

$$P(Y=1) = P(Y=2) = P(Y=3) = P(Y=4) = P(Y=5) = \frac{4}{25},$$
$$P(Y=6) = \frac{5}{25}.$$

R:llä nopanheittojen simulointi ja havaittujen frekvenssien laskeminen sujuvat seuraavasti.

```
n <- 2000
p <- c(4, 4, 4, 4, 4, 5)
p <- p/sum(p)
y <- sample(1:6, size = n, replace = TRUE, prob = p)
obs <- table(y)
```

Testaan nyt eräässä tällaisessa simuloinnissa saatuja frekvenssejä, kun (epätoden) nollahypoteesin mukaan kaikilla silmäluvuilla on sama todennäköisyys 1/6. Havaitut frekvenssit, odotetut frekvenssit sekä suureet $(O_i - E_i)^2/E_i$ ovat

| silmäluku | 1 | 2 | 3 | 4 | 5 | 6 | summa |
|---|---|---|---|---|---|---|---|
| havaittu $O_i$ | 311 | 318 | 306 | 342 | 316 | 407 | 2000 |
| odotettu $E_i$ | 333.3 | 333.3 | 333.3 | 333.3 | 333.3 | 333.3 | 2000 |
| $(O_i - E_i)^2/E_i$ | 1.496 | 0.705 | 2.241 | 0.225 | 0.901 | 16.280 | 21.85 |

Allaolevassa koodissa lasketaan Pearsonin $X^2$-testisuureen arvo, $\chi^2_{k-1}$-jakauman kriittinen arvo sekä testin $p$-arvo.

```
obs <- c(`1` = 311, `2` = 318, `3` = 306, `4` = 342, `5` = 316,
    `6` = 407)
n <- sum(obs)
alpha <- 0.05
p0 <- c(1, 1, 1, 1, 1, 1)/6
expected <- n * p0
print(x2 <- sum((obs - expected)^2/expected))

## [1] 21.85

nu <- length(p0) - 1
print(crit <- qchisq(alpha, df = nu, lower = FALSE))
```

```
## [1] 11.07

print(p.value.x2 <- pchisq(x2, df = nu, lower = FALSE))

## [1] 0.0005591
```

Nollahypoteesi hylätään merkitsevyystasolla $\alpha = 0.05$. Sama testi saadaan suoritettua myös soveltamalla R:n funktiota `chisq.test`.

```
chisq.test(obs)

##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 21.85, df = 5, p-value = 0.0005591
```

△

**Esimerkki 8.2** Tarkistetaan seuraavaksi, miten käy kun Pearsonin testisuure lasketaan binomikokeessa, jossa onnistumisten lukumäärä $K$ noudattaa binomijakaumaa $\mathrm{Bin}(n, p)$. Tällöin saadaan taulukko

| luokka | onnistuminen | epäonnistuminen | yhteensä |
|---|---|---|---|
| havaittu | $K$ | $n - K$ | $n$ |
| odotettu | $np$ | $n(1-p)$ | $n$ |

Tällöin

$$\begin{aligned} X^2 &= \frac{(K-np)^2}{np} + \frac{[(n-K) - n(1-p)]^2}{n(1-p)} \\ &= \frac{(K-np)^2}{np(1-p)}. \end{aligned}$$

Pearsonin testisuureen teorian mukaan tämän satunnaismuuttujan pitäisi likimain noudattaa jakaumaa $\chi^2_1$. Toisaalta binomijakauman normaalijakauma-approksimaation mukaan satunnaismuuttuja

$$\frac{K - np}{\sqrt{np(1-p)}}$$

noudattaa suurella otoskoolla $n$ osapuilleen standardinormaalijakaumaa $N(0, 1)$, ja Pearsonin testisuure $X^2$ on yhtä kuin tämän satunnaismuuttujan neliö. Todennäköisyyslaskennasta tiedetään, että mikäli $Z \sim N(0, 1)$, niin

$$Z^2 \sim \chi^2_1.$$

Nämä tulokset antavat tuntumaa siihen, milloin khiin neliön jakauma-approksimaatio toimii hyvin ja milloin taas huonosti. △

Pearsonin $X^2$-testisuureeseen perustuva testi on likimääräinen, sillä se perustuu likimääräiseen suuren otoskoon jakaumatulokseen. Milloin tämä approksimaatio on tarpeeksi hyvä? Kirjallisuudessa löytyy tähän tilanteeseen erilaisia suosituksia. Testiä sovelletaan huolta vailla esim. silloin, jos kaikille luokille niiden odotetut frekvenssit ovat vähintään viisi. Jos joidenkin luokkien odotetut frekvenssit ovat liian pieniä, niin sitten kyseisiä luokkia voidaan yhdistää keskenään ennen testin soveltamista.

Yhteensopivuustestien voima kasvaa otoskoon kasvaessa. Jos nollahypoteesi ei pidä tarkasti paikkaansa, niin kyllin suurella otoskoolla se jossakin vaiheessa hylätään, vaikka poikkeama olisi niin pieni, että sillä ei ole käytännön kannalta merkitystä.

Usein yhteensopivuustestissä nollahypoteesi on yhdistetty, ja se voidaan esittää kaavalla

$$H_0 : p_1 = p_1(\theta), \ldots, p_k = p_k(\theta), \qquad \text{jollekin } \theta \in \Theta_0. \tag{8.7}$$

Tällöin odotettuja frekvenssejä ei saada laskettua ennen kuin parametrin $\theta$ arvo on estimoitu. Mikäli $\theta$ estimoidaan havaituista frekvensseistä suurimman uskottavuuden menetelmällä, ja estimaatti on $\hat{\theta}$, niin tällöin testisuureena voidaan käyttää tuttua Pearsonin testisuuretta,

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Uusi asia on se, että odotetut frekvenssit pitää laskea käyttämällä parametrin tilalla sen SU-estimaattia

$$E_i = n\, p_i(\hat{\theta}), \qquad i = 1, \ldots, k. \tag{8.8}$$

Jos parametrilla $\theta$ on $d$ vapaata komponenttia (ts. komponenttia, joita mikään sidosehto ei kytke toisiinsa) niin tällöin testisuureella (satunnaismuuttujaksi ymmärrettynä) on suurella otoskoolla osapuilleen $\chi^2$-jakauma vapausasteluvulla

$$\nu = k - 1 - d. \tag{8.9}$$

Tämän asian perusteli ensimmäisenä Fisher 1920-luvulla. Hän osoitti samalla, että K. Pearson oli tehnyt omissa laskelmissaan virheen vapausasteluvun kohdalla.

Ennenkuin SU-menetelmää voidaan käyttää, pitää tietenkin pystyä kirjoittamaan uskottavuusfunktio. Yhden havaintosatunnaismuuttujan $Y_h$ pistetodennäköisyysfunktio voidaan ilmaista kaavalla

$$\begin{aligned} P(Y_h = y_h) &= \begin{cases} p_1 & \text{jos } y_h = 1, \\ p_2 & \text{jos } y_h = 2, \\ \vdots & \\ p_k & \text{jos } y_h = k. \end{cases} \\ &= p_1^{1(y_h=1)}\, p_2^{1(y_h=2)} \cdots p_k^{1(y_h=k)} \end{aligned}$$

Tässä $1(y_h = i)$ on osoitinmuuttuja sille, että $y_h$:n arvo on $i$, eli

$$1(y_h = i) = \begin{cases} 1 & \text{mikäli } y_h = i, \\ 0 & \text{muuten.} \end{cases}$$

Havaintoja $y_1, \ldots, y_n$ vastaava uskottavuusfunktio on

$$L(p_1, \ldots, p_{k-1}) = \prod_{h=1}^{n} p_1^{1(y_h=1)} p_2^{1(y_h=2)} \cdots p_k^{1(y_h=k)}$$

Kun luokkien todennäköisyyksien $p_i$ potenssit yhdistetään, uskottavuusfunktiolle saadaan lauseke

$$L(p_1, \ldots, p_{k-1}) = p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \tag{8.10}$$

jossa $n_i$ on niiden indeksien $h$ lukumäärä, joille $y_h = i$, eli $n_i$ on luokan $i$ havaittu frekvenssi. Jos luokkien todennäköisyydet $p_i$ riippuvat parametrin $\theta$ arvosta, niin uskottavuusfunktioksi saadaan

$$L(\theta) = p_1(\theta)^{n_1} p_2(\theta)^{n_2} \ldots p_k(\theta)^{n_k}, \qquad \theta \in \Theta.$$

**Esimerkki 8.3** Kalbfleish [2, Esimerkki 11.2.2] esittää aineiston, jossa on etsitty hiukkaslaskurilla alfahiukkasia. Kokeessa on toistettu 2608 kertaa tietyn pituinen mittaus, jossa on laskettu laskuriin osuvien alfahiukkasten lukumäärä. Sitten on taulukoitu niiden mittausten lukumäärä, joissa laskuri on havainnut $i$ kappaletta alfahiukkasia, kun $i = 0, 1, \ldots, 9$, mutta kaikki tapaukset joissa on havaittu $i \geq 10$ alfahiukkasta on yhdistetty yhdeksi luokaksi. Havaitut frekvenssit ovat seuraavassa taulukossa.

| luokka | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $\geq 10$ | yht. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| havaittu | 57 | 203 | 383 | 525 | 532 | 408 | 273 | 139 | 45 | 27 | 16 | 2608 |

Kokeessa tutkitaan, noudattaako alfahiukkasten lukumäärä Poissonin jakaumaa. Poissonin jakauma on diskreetti todennäköisyysjakauma, jonka pistetodennäköisyysfunktio on

$$g(x; \theta) = \mathrm{e}^{-\theta} \frac{\theta^x}{x!}, \qquad x = 0, 1, 2, \ldots,$$

jossa parametri $\theta > 0$ on jakauman odotusarvo. Tällöin luokat $0, \ldots, 9$ vastaavat havaittujen alfahiukkasten lukumääriä, ja niiden todennäköisyydet ovat

$$p_i(\theta) = g(i; \theta), \qquad i = 0, 1, \ldots, 9.$$

Viimeinen luokka vastaa sitä tapahtumaa, että laskuri havaitsee vähintään 10 alfahiukkasta, ja sen todennäköisyys on

$$p_{10}(\theta) = 1 - \sum_{i=0}^{9} g(i; \theta).$$

Luokkien todennäköisyyksiä ei saada laskettua ennen kuin Poissonin jakauman parametri on estimoitu. SU-estimaatiksi saadaan (tietokoneen avulla) $\hat{\theta} = 3.8703$. Kun tämä sijoitetaan Poissonin jakauman pistetodennäköisyysfunktioon, saadaan Pearsonin testisuure $X^2$ laskettua.

| luokka | havaittu | odotettu | $(O_i - E_i)^2/E_i$ |
|---|---|---|---|
| 0 | 57 | 54.38 | 0.13 |
| 1 | 203 | 210.47 | 0.27 |
| 2 | 383 | 407.30 | 1.45 |
| 3 | 525 | 525.46 | 0.00 |
| 4 | 532 | 508.42 | 1.09 |
| 5 | 408 | 393.55 | 0.53 |
| 6 | 273 | 253.86 | 1.44 |
| 7 | 139 | 140.36 | 0.01 |
| 8 | 45 | 67.90 | 7.73 |
| 9 | 27 | 29.20 | 0.17 |
| $\geq 10$ | 16 | 17.08 | 0.07 |
| summa | 2608 | 2608 | 12.88 |

Arvoa $X^2 = 12.88$ verrataan $\chi^2_\nu$-jakaumaan, kun vapausasteluku on

$$\nu = k - 1 - d = 11 - 1 - 1 = 9.$$

Testin $p$-arvoksi saadaan $p = 0.17$. Tämä ei ole erityisen pieni: jos Poisson-malli pitää paikkaansa, niin 17 %:n todennäköisyydellä saadaan testisuureelle arvoja joiden mukaan havaitut ja odotetut frekvenssit poikkeavat ainakin näin paljon toisistaan. Nollahypoteesi jää voimaan, jos käytetään esim. 5 %:n merkitsevyystasoa. △

## 8.2 Riippumattomuuden testaaminen kontingenssitaulukossa

Tarkastellaan $n$ otosyksikköä, joista kustakin mitataan kaksi diskreettiä ominaisuutta $(x_h, y_h)$, jossa $h = 1, \ldots, n$. Ominaisuus $x_h$ saa yhden arvoista $1, \ldots, r$ ja ominaisuus $y_h$ yhden arvoista $1, \ldots, c$.

Tilastollinen malli koostuu $n$ riippumattomasta ja samoin jakautuneesta satunnaismuuttujaparista $(X_h, Y_h)$, joille

$$p_{ij} = P(X_h = i, Y_h = j), \qquad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

Tilanne on muuten samanlainen kuin edellisessä jaksossa, mutta nyt luokkia indeksoidaan kahdella indeksillä $i$ ja $j$ eikä enää yhdellä indeksillä. Luokkia on $rc$, joten vapaita parametreja $p_{ij}$ on $rc - 1$, mikäli niitä ei rajoiteta lisäämällä malliin oletuksia.

Havaitut frekvenssit ovat

$$n_{ij}, \qquad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

ja ne voidaan esittää taulukkona, jossa $i$ indeksoi vaakarivejä ja $j$ sarakkeita. Tälläistä taulukkoa kutsutaan *kontingenssitaulukoksi* (engl. *contingency table*). Taulukon rivin $i$ summaa merkitään $n_{i\bullet}$ ja sarakkeen $j$ summaa $n_{\bullet j}$, ts. alain-

**Taulukko 8.1** Kontingenssitaulukko, jolla on $r$ riviä ja $c$ saraketta.

| | | | | |
|---|---|---|---|---|
| $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1c}$ | $n_{1\bullet}$ |
| $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2c}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rc}$ | $n_{r\bullet}$ |
| $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet c}$ | $n$ |

deksi piste tarkoittaa summaamista kyseisen indeksin yli, eli

$$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}, \qquad i = 1, \ldots, r$$

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}, \qquad j = 1, \ldots, c.$$

Testattavan nollahypoteesin mukaan satunnaismuuttujat $X_h$ ja $Y_h$ ovat riippumattomat, mikä tarkoittaa sitä, että kaikilla $i$ ja $j$ niiden yhteisjakauman pistetodennäköisyysfunktio saadaan kertomalla keskenään vastaavat reunajakaumien pistetodennäköisyydet, eli

$$p_{ij} = P(X_h = i, Y_h = j) = P(X_h = i)\, P(Y_h = j) = \gamma_i\, \delta_j, \quad \text{missä}$$
$$\gamma_i = P(X_h = i), \qquad \delta_j = P(Y_h = j).$$

Reunajakaumien pistetodennäköisyydet $\gamma_1, \ldots, \gamma_r$ ja $\delta_1, \ldots, \delta_c$ ovat tuntemattomia parametreja. Nollahypoteesi voidaan ilmaista myös sanomalla, että rivi- ja sarakeluokittelut ovat riippumattomia.

Jaksossa 8.5 osoitetaan, että nollahypoteesin vallitessa suurimman uskottavuuden estimaatit ovat

$$\hat{\gamma}_i = \frac{n_{i\bullet}}{n}, \qquad i = 1, \ldots, r \tag{8.11}$$

$$\hat{\delta}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \ldots, c. \tag{8.12}$$

Jälleen kerran todennäköisyysparametrien suurimman uskottavuuden estimaatit ovat vastaavat suhteelliset frekvenssit.

Kun nollahypoteesi pitää paikkansa, niin tuntemattomia vapaita parametreja on

$$d = r - 1 + c - 1$$

kappaletta, sillä molemmat reunatodennäköisyysfunktiot summautuvat ykköseksi. Khiin neliön testissä vapausasteiden lukumääräksi saadaan kaavan (8.9) mukaan

$$\nu = rc - 1 - (r - 1 + c - 1) = (r - 1)(c - 1).$$

Havaitut frekvenssit ovat $O_{ij} = n_{ij}$, ja nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_{ij} = n\, \hat{\gamma}_i\, \hat{\delta}_j = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$$

Pearsonin $\chi^2$-testisuure on

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Koon $0 < \alpha < 1$ testissä testisuuretta verrataan $\chi^2_\nu$ jakauman $\alpha$-yläkvantiiliin, jossa vapausasteluku

$$\nu = (r-1)(c-1).$$

**Esimerkki 8.4** (ABO-veriryhmien geneettinen perusta [1, Jakso 4.5]) Ihmisiä luokitellaan eri veriryhmiin veren ominaisuuksien perusteella. Tunnetuin veriryhmäjärjestelmä on ns. ABO-veriryhmäjärjestelmä, jossa kunkin yksilön veriryhmä on joko A, B, AB tai O. Veriryhmä määritetään tarkistamalla, onko henkilön veressä antigeeniä A tai antigeeniä B. Veriryhmät nimetään tuloksen perusteella seuraavan taulukon mukaisesti

| | Ei B | On B |
|---|---|---|
| Ei A | O | B |
| On A | A | AB |

Vielä 1920-luvulla oli epäselvää, minkälaisesta geneettisestä mekanismista ABO-veriryhmät määräytyvät. Yksi mahdollinen selitys oli kahden riippumattoman lokuksen malli, jossa lokuksen yksi alleeli määrää, onko veressä antigeeniä A vai ei, ja lokuksen kaksi alleeli määrää, onko veressä antigeeniä B vai ei. Jos kahden riippumattoman lokuksen malli on tosi ja jos otamme populaatiosta satunnaisotoksen, niin tällöin veriryhmistä muodostetussa kontingenssitaulukossa rivi- ja sarakeluokittelut ovat riippumattomia. Englannissa 1930-luvulla tehdystä otoksesta saatiin seuraava veriryhmien kontingenssitaulukko

| | Ei B | On B |
|---|---|---|
| Ei A | 202 | 35 |
| On A | 179 | 6 |

Tästä taulukosta laskettuna Pearsonin $\chi^2$ testisuureen arvo on $X^2 = 15.73$. Tätä verrataan $\chi^2_1$-jakauman $\alpha$-yläkvantiiliin. Esim. jos merkitsevyystaso on $\alpha = 0.05$, niin tämä kriittinen arvo on 3.84, joten kahden lokuksen malli tulee testissä hylättyä. R:llä testisuureiden arvot voidaan laskea seuraavasti.

```
print(bloodgroups <- matrix(c(202, 179, 35, 6), 2, 2))

##      [,1] [,2]
## [1,]  202   35
## [2,]  179    6

print(n <- sum(bloodgroups))

## [1] 422

print(gamma.hat <- rowSums(bloodgroups)/n)

## [1] 0.5616 0.4384

print(delta.hat <- colSums(bloodgroups)/n)

## [1] 0.90284 0.09716

observed <- bloodgroups
print(expected <- n * (gamma.hat %o% delta.hat))
```

```
##      [,1]  [,2]
## [1,]  214 23.03
## [2,]  167 17.97

print(x2 <- sum((observed - expected)^2/expected))

## [1] 15.73

nu <- (2 - 1) * (2 - 1)
print(crit <- qchisq(0.05, df = nu, lower = FALSE))

## [1] 3.841

print(p.x2 <- pchisq(x2, df = nu, lower = FALSE))

## [1] 7.298e-05
```

Testin voi suorittaa myös funktiolla `chisq.test`. Tämä funktio käyttää oletusarvoisesti ns. jatkuvuuskorjausta. Alla olevassa koodissa pyydetään erikseen olemaan käyttämättä jatkuvuuskorjausta, jotta tuloksia voitaisiin suoraan verrata itse tehtyjen laskujen kanssa.

```
chisq.test(bloodgroups, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  bloodgroups
## X-squared = 15.73, df = 1, p-value = 7.298e-05
```

Nykyään tiedetään, että ABO-veriryhmä määräytyy yhdestä lokuksesta, jolla voi olla kolme alleelia A, B tai O, joista A ja B ovat dominoivia ja O resessiivinen. Nollahypoteesin hylkääminen on oikea päätös. △

## 8.3 Homogeenisuuden testaaminen

Kuten edellisessä jaksossa, nytkin oletetaan, että otosyksiköillä on kaksi diskreettiä ominaisuutta $x$ ja $y$, ja $x$:n mahdolliset arvot ovat $1, \ldots, r$ ja $y$:n mahdolliset arvot $1, \ldots, c$. Populaatio jaetaan ominaisuuden $x$ arvojen määräämiin ositteisiin siten, että ositteessa $i$ ominaisuuden $x$ arvo on $i$. Kustakin osasta tehdään riippumaton kokoa $n_i$ oleva otos

$$Y_{ih}, \qquad h = 1, \ldots, n_i.$$

Tavoitteena on testata, ovatko ositteiden jakaumat samat eli ovatko ositteet homogeenisia. Nollahypoteesi on

$$H_0 : p_{ij} = \pi_j, \qquad \text{kaikilla } i = 1, \ldots, r \text{ ja } j = 1, \ldots, c, \tag{8.13}$$

missä

$$p_{ij} = P(Y_{ih} = j),$$

ja todennäköisyydet $(\pi_1, \ldots, \pi_c)$ ovat tuntemattomia.

Todennäköisyyksien $(\pi_i)$ suurimman uskottavuuden estimaateiksi saadaan

$$\hat{\pi}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \ldots, c. \tag{8.14}$$

Testissä muodostetaan taulukko, jossa vaakariville $i$ tulee ositteesta $i$ lasketut frekvenssit, ja vaakarivin $i$ frekvenssien summa $n_{i\bullet}$ on ositteesta $i$ tehdyn otoksen koko $n_i$. Osoittautuu, että tämän jälkeen testaus voidaan tehdä aivan samalla tavalla kuin edellisessä jaksossa. Pearsonin testisuureetta verrataan $\chi^2_\nu$-jakaumaan, jossa vapausasteluku on

$$\nu = (r-1)(c-1).$$

## 8.4 Uskottavuusosamäärän testisuure

Tämän luvun tilanteissa käytetään usein myös muita testisuureita kuin Pearsonin testisuureetta. Merkitään nyt tilastollisen mallin uskottavuusfunktiota $L(\boldsymbol{\theta}; \mathbf{Y})$, sen logaritmista uskottavuusfunktiota $\ell(\boldsymbol{\theta}; \mathbf{Y})$ ja suurimman uskottavuuden estimaattoria symbolilla $\hat{\boldsymbol{\theta}}(\mathbf{Y})$. Suurimman uskottavuuden estimaattorien asymptoottisen teorian perusteella tiedetään, että jos havaintosatunnaisvektorilla $\mathbf{Y}$ on parametria $\boldsymbol{\theta}$ vastaava jakauma, niin tällöin *uskottavuusosamäärän testisuureella* (engl. *likelihood ratio statistic*)

$$W = 2[\ell(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y}) - \ell(\boldsymbol{\theta}; \mathbf{Y})] = 2 \log \frac{L(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y})}{L(\boldsymbol{\theta}; \mathbf{Y})}$$

on (tiettyjen oletusten vallitessa) $\chi^2$-jakauma, jossa vapausasteluku on yhtä kuin mallin vapaiden parametrien lukumäärä.

Tarkastelemme ensin jakson 8.1 tilastollista mallia silloin, kun luokkien todennäköisyydet $p_1, \ldots, p_{k-1}$ ovat vapaita parametreja. Uskottavuusosamäärän testisuureessa tarvittavat log-uskottavuusarvot ovat kaavan (8.10) mukaan

$$\ell(\hat{p}_1, \ldots, \hat{p}_{k-1}; \mathbf{y}) = \sum_{i=1}^{k} n_i \, \log \hat{p}_i$$

$$\ell(p_1, \ldots, p_{k-1}; \mathbf{y}) = \sum_{i=1}^{k} n_i \, \log p_i$$

Näytämme seuraavassa jaksossa, että SU-estimaatit ovat vastaavat suhteelliset frekvenssit

$$\hat{p}_1 = \frac{n_1}{n}, \ldots, \hat{p}_{k-1} = \frac{n_{k-1}}{n}, \hat{p}_k = \frac{n_k}{n},$$

joten uskottavuusosamäärän testisuure on

$$W = 2 \sum_{i=1}^{k} n_i \, \log \frac{\hat{p}_i}{p_i} = 2 \sum_{i=1}^{k} n_i \, \log \frac{n_i}{n \, p_i} \tag{8.15}$$

$$= 2 \sum_{i=1}^{k} O_i \, \log \frac{O_i}{E_i}, \tag{8.16}$$

Jos jokin $n_i$ tai $O_i$ on nolla, niin tässä kaavassa pitää tulkita $0 \log 0 = 0$. Suurella otoskoolla testisuuretta vastaavalla satunnaismuuttujalla on osapuilleen $\chi^2_{k-1}$-jakauma.

Myös muissa tämän kappaleen tilanteissa voidaan myös käyttää Pearsonin testisuureen sijasta uskottavuusosamäärän testisuuretta, mutta tällöin tarvitaan sisäkkäisten mallien vertailuun tarkoitettua uskottavuusosamäärän testisuureetta [1, kaava (4.38)] Se on kaikissa tämän kappaleen tilanteissa muotoa

$$W = 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i},$$

tai kaksiulotteisille taulukoille muotoa

$$W = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log \frac{O_{ij}}{E_{ij}},$$

jossa odotetut frekvenssit lasketaan samoin kuin Pearsonin testisuureelle. Uskottavuusosamäärän testisuureetta verrataan täsmälleen samaan $\chi^2$-jakaumaan kuin Pearsonin testisuuretta.

## 8.5 Suurimman uskottavuuden estimaatit

Tässä jaksossa johdamme tässä kappaleessa ilmoitetut suurimman uskottavuuden estimaattien kaavat. Kaavat olisi mahdollista johtaa monella menetelmällä. Voisimme eliminoida yhden todennäköisyysparametereista käyttämällä sitä tietoa, että ne summautuvat ykköseksi. Toinen mahdollisuus olisi käyttää Lagrangen keinoa rajoitteellisten optimointitehtävien ratkaisemiseksi. Tässä jaksossa tulokset kuitenkin johdetaan käyttämällä juuri tähän tilanteeseen sopivaa epäyhtälöä. Tällä konstilla johdoista tulee paljon yksinkertaisempia kuin yleisemmillä keinoilla.

Johdamme ensin parametreille $p_1, \ldots, p_k$ eli luokkien todennäköisyyksille suurimman uskottavuuden estimaatit jakson 8.1 tilastollisessa mallissa siinä tapauksessa, kun $k-1$ luokan todennäköisyydet $p_1, \ldots, p_{k-1}$ ovat vapaita parametreja. Otamme johdossa huomioon, että parametreilla on lineaarinen rajoite

$$p_1 + \cdots + p_k = 1$$

ja että havainnoilla $n_i$ on rajoite

$$n_1 + \cdots + n_k = n.$$

Käytämme hyväksi aputulosta, joka sanoo, että luonnollisen logaritmin $\log(x)$ kuvaaja jää pisteeseen $x = 1$ piirretyn tangenttinsa alapuolelle.

$$\log(x) \leq x - 1, \qquad \text{kaikilla } x > 0. \tag{8.17}$$

Yhtäsuuruus saavutetaan ainoastaan pisteessä $x = 1$. Tämän väitteen voi tarkistaa helposti analyysin keinoilla.

Estimaatien kaavat nähdään helposti seuraavan lauseen avulla.

**Lause 8.1.** *Jos $k$ ei-negatiivista lukua $n_i \geq 0$ summautuvat luvuksi $n > 0$, eli*

$$\sum_{i=1}^{k} n_i = n,$$

*niin tällöin mille tahansa luvuille $p_i \geq 0$ jotka summautuvat ykköseksi pätee*

$$\sum_{i=1}^{k} n_i \log p_i \leq \sum_{i=1}^{k} n_i \log \frac{n_i}{n},$$

*missä käytämme tarvittaessa sopimusta $0 \log 0 = 0$.*

*Todistus.* Esitän todistuksen vain siinä tapauksessa, jossa kaikki $n_i > 0$. Yleisen tapauksen saa todistettua helposti samaan tapaan. Väitetyn epäyhtälön vasemman ja oikean puolen erotus on

$$\begin{aligned}\sum \left( n_i \log p_i - n_i \log \frac{n_i}{n} \right) &= \sum n_i \log \frac{p_i}{n_i/n} \leq \sum n_i \left( \frac{p_i}{n_i/n} - 1 \right) \\ &= n \sum p_i - \sum n_i = n - n = 0,\end{aligned}$$

missä sovelsimme epäyhtälöä (8.17). □

Suurimman uskottavuuden estimaatit ovat siis vastaavat suhteelliset frekvenssit, eli

$$\hat{p}_i = \frac{n_i}{n}, \qquad i = 1, \ldots, k. \tag{8.18}$$

Jos testataan riippumattomuutta kontingenssitaulukossa, niin logaritminen uskottavuusfunktio on nollahypoteesin $p_{ij} = \gamma_i \, \delta_j$ vallitessa

$$\begin{aligned}\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log p_{ij} &= \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log(\gamma_i \, \delta_j) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, [\log \gamma_i + \log \delta_j] \\ &= \sum_{i=1}^{r} \log \gamma_i \sum_{j=1}^{r} n_{ij} + \sum_{j=1}^{c} \log \delta_j \sum_{i=1}^{r} n_{ij} \\ &= \sum_{i=1}^{r} n_{i\bullet} \log \gamma_i + \sum_{j=1}^{c} n_{\bullet j} \log \delta_j \\ &\leq \sum_{i=1}^{r} n_{i\bullet} \log \frac{n_{i\bullet}}{n} + \sum_{j=1}^{c} n_{\bullet j} \log \frac{n_{\bullet j}}{n},\end{aligned}$$

mistä nähdään, että suurimman uskottavuuden estimaateilla on kaavat (8.11) ja (8.12).

Homogeenisuuden testauksessa uskottavuusfunktio on nollahypoteesin $p_{ij} = \pi_j$ vallitessa

$$L(\pi_1, \ldots, \pi_{c-1}) = \prod_{i=1}^{r} \pi_1^{n_{i1}} \, \pi_2^{n_{i2}} \ldots \pi_c^{n_{ic}} = \pi_1^{n_{\bullet 1}} \, \pi_2^{n_{\bullet 2}} \ldots \pi_c^{n_{\bullet c}},$$

joten suurimman uskottavuuden estimaattien kaavat (8.14) saadaan suoraan lauseesta 8.1, kun ensin siirrytään tarkastelemaan uskottavuusfunktion logaritmia.

# Kirjallisuutta

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

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

# Luku 9

# Lineaarinen regressio

## 9.1 Johdanto

Oletamme, että havaintoaineisto koostuu kahden muuttujan $x$ ja $y$ arvoista

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

siten että pari $(x_i, y_i)$ on mitattu havaintoyksiköstä $i$. Emme aluksi lainkaan aseta tilastollista mallia, vaan tarkastelemme tilannetta heuristisesti.

Päämääränä regressiomalleissa on kuvailla, kuinka *selittävä muuttuja* $x$ (engl. *explanatory variable*, *independent variable*) vaikuttaa *selitettävään muuttujaan* eli *vasteeseen* $y$ (engl. *dependent variable*, *response*). Yleisemmässä tilanteessa selittäviä muuttujia voisi olla usampia, mutta tässä kappaleessa selittäjiä on vain yksi.

Ajattelemme, että $x$ vaikuttaa $y$:n arvoon osapuilleen kaavan

$$y = f(x)$$

mukaisesti, jossa $f$ on funktio, jonka muoto on ainakin osittain tuntematon. Tarkemmin sanoen funktion arvo $f(x)$ riippuu $x$:n arvon lisäksi tuntemattomien parametrien arvoista vaikka tätä ei olla merkinnöissä huomioitu. Erilaisten virhelähteiden takia pisteet $(x_i, y_i)$ eivät kuitenkaan tarkalleen asetu funktion $f$ kuvaajalle millään parametrin arvolla, minkä takia tyydymme malliin

$$y = f(x) + \epsilon,$$

jossa muuttuja $\epsilon$ tarkoittaa virhettä. Funktion $f$ muoto (ts. parametrien arvot) pitäisi määrittää havaintojen perusteella.

Tämän yleisen rakennemallin $f(x)$ sijasta tarkastelemme lineaarista lauseketta

$$\alpha + \beta\, x,$$

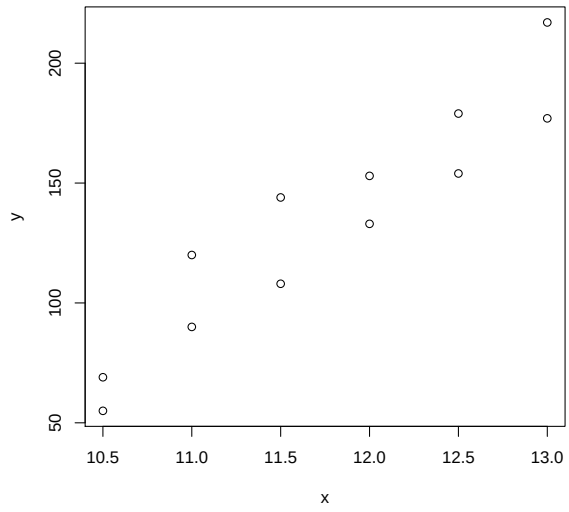
jonka kuvaaja on suora. Jos $x$-arvojen vaihteluväli on pieni, niin melkein mitä tahansa funktiota $f(x)$ voidaan approksimoida tällaisella lineaarisella funktiolla.

Lineaarisessa rakennemallissa $\alpha + \beta\, x$ parametri $\alpha$ on suoran vakiotermi (engl. *intercept*) ja parametri $\beta$ on suoran kulmakerroin (engl. *slope*). Ne ovat tuntemattomia parametreja, joiden arvot pitää määrittää havaintojen perusteella. Lähdemme seuraavaksi etsimään tälle tehtävälle toimivaa ratkaisua.

**Taulukko 9.1** Kuminauhan lentomatkoja $y$ eri venytyksen $x$ arvoilla.

| x | y | x | y |
|---|---|---|---|
| 11 | 120 | 10.5 | 69 |
| 12 | 153 | 10.5 | 55 |
| 13 | 217 | 11.5 | 108 |
| 13 | 177 | 11.5 | 144 |
| 12 | 133 | 12.5 | 154 |
| 11 | 90 | 12.5 | 179 |

**Kuva 9.1** Kuminauhan lentomatkoja $y$ eri venytyksen $x$ arvoilla.

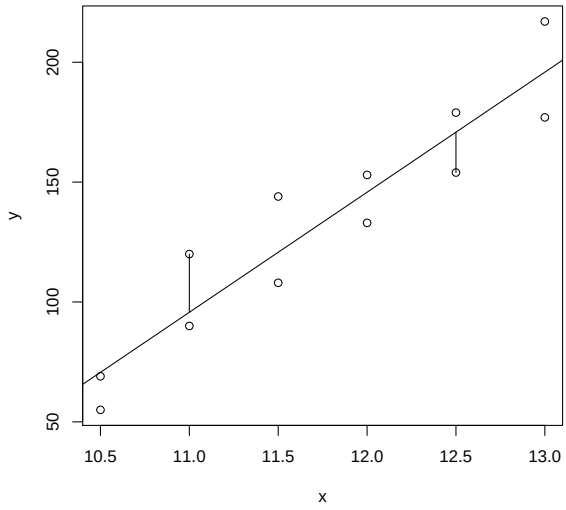



Analysoimme tässä kappaleessa kurssin luennoitsijan kotonaan mittaamaa aineistoa, jossa tutkittiin, miten pitkälle kuminauha lentää, jos sitä ensin venytetään ja sitten sen toinen pää päästetään vapaaksi. Kuminauhaa venytettiin siten, että sen toista päätä pingoitettiin viivottimen nollapisteen puoleiseen päähän ja venytystä kontrolloitiin asettamalla toinen pää tiettyyn kohtaan $x$ (cm) viivoittimen asteikolla. Viivoitin ja kuminauha pidettiin vaakasuorassa 50 cm:n korkeudessa lattialta (pöydän avulla), ja lentomatka $y$ (cm) mitattiin katsomalla, miten kauas kuminauhan kauimmainen kohta päätyi siitä lattian pisteestä, jonka yläpuolella viivottimen nollapisteen puoleinen pää oli. Mittaukset ovat taulukossa 9.1, ja kuvassa 9.1 ne esitetään hajontakuvana.

## 9.2 Suoran sovittaminen pienimmän neliösumman menetelmällä

Emme vieläkään aseta tilastollista mallia, vaan tarkastelemme edelleen tehtävää heuristisesti. Parametrit eli kertoimet $\alpha$ ja $\beta$ voidaan määrittää sellaisella taval-

**Kuva 9.2** Kuminauha-aineistoon sovitettu pienimmän neliösumman suora $a + bx$, jossa $a = -455.3$ ja $b = 50.09$.




la, jossa yritetään saada vasteet $y_i$ sekä niitä vastaavat sovitteet tai ennusteet $\alpha+\beta x_i$ mahdollisimman samankaltaisiksi jonkin kriteerin mielessä. Tässä yhteydessä samankaltaisuutta olisi mahdollista mitata erilaisilla tavoilla, ja erilaiset tavat johtaisivat erilaisin parametriestimaatteihin.

Käytännössä tärkein kriteeri on virheiden neliöiden summa

$$\mathrm{SS}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{9.1}$$

Tässä $|y_i - \alpha - \beta x_i|$ on pisteen $(x_i, y_i)$ $y$-akselin suunnassa mitattu etäisyys suorasta $\alpha + \beta x$. Kriteerissä $\mathrm{SS}(\alpha, \beta)$ nämä $y$-akselin suuntaiset etäisyydet neliöidään ja sitten ne summataan yhteen. Tämän kriteerin mielessä paras suora eli pienimmän neliösumman suora on se suora, jota vastaavat kertoimet $\alpha$ ja $\beta$ minimoivat virheiden neliöiden summan (9.1). Voimme myös sanoa, että sovitamme suoran aineistoon käyttämällä pienimmän neliösumman menetelmää (engl. *method of least squares*) eli PNS-menetelmää. Kuvassa 9.2 on piirretty kuminauha-aineisto sekä siihen sovitettu PNS-suora. Lisäksi kahdelle $(x, y)$-pisteelle on piirretty sen $y$-akselin suuntaan mitattu etäisyys PNS-suorasta.

Kriteeri (9.1) on tärkeä toisaalta sen takia, että se johtaa yksinkertaisiin kaavoihin ja toisaalta sen takia, että sen käyttö voidaan perustellä myöhemmin esitettävällä yksinkertaisella tilastollisella mallilla sekä suurimman uskottavuuden periaattella.

Todistamme jakson lopussa, että PNS-menetelmä valitsee kulmakertoimelle $\beta$ arvon

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{9.2}$$

Tässä tuttuun tapaan $\bar{x}$ ja $\bar{y}$ tarkoittavat keskiarvoja

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Jotta $b$ olisi hyvin määritelty, täytyy olettaa, että kaavan (9.2) nimittäjä on aidosti positiivinen. Tämä on voimassa, mikäli jonossa $x_1, \ldots, x_n$ on vähintään kaksi erisuurta arvoa, ja tämän ehdon oletamme jatkossa. Kun kulmakerroin on laskettu, niin PNS-menetelmä antaa vakion $\alpha$ arvoksi

$$\hat{\alpha} = a = \bar{y} - b\bar{x}. \tag{9.3}$$

Huomaa että tuntemattomia parametreja (tai kertoimia) merkitään kreikkalaisilla kirjaimilla, ja niiden estimaatteja voidaan merkitä joko laittamalla hattu parametrin päälle tai sitten käyttämällä kreikkalaista kirjainta vastaavaa latinalaista kirjainta.

PNS-suoran kulmakertoimelle voidaan antaa tulkinta otoskovarianssin ja otosvarianssin avulla. Määrittelemme vektoreista

$$\mathbf{x} = (x_1, \ldots, x_n), \qquad \mathbf{y} = (y_1, \ldots, y_n).$$

lasketun otoskovarianssin $s(\mathbf{x}, \mathbf{y})$ kaavalla

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{9.4}$$

Otoskovarianssissa käytetään jakajana lukua $n-1$ sen takia, että näin saadaan populaatiokovarianssin harhaton estimaattori. Jos $X$ ja $Y$ ovat satunnaismuuttujia, niin niiden kovarianssi (eli populaatiokovarianssi) $\operatorname{cov}(X, Y)$ määritellään odotusarvona

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)]. \tag{9.5}$$

Jos $(X_1, Y_1), \ldots, (X_n, Y_n)$ on satunnaisotos satunnaismuuttujaparin $(X, Y)$ jakaumasta, niin helpohkoilla laskuilla nähdään, että

$$Es(\mathbf{X}, \mathbf{Y}) = \operatorname{cov}(X, Y),$$

missä tietenkin $\mathbf{X} = (X_1, \ldots, X_n)$ ja $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

PNS-suoran kulmakerroin (9.2) saadaan lausuttua otoskovarianssin (9.4) avulla kaavalla

$$b = \frac{s(\mathbf{x}, \mathbf{y})}{s(\mathbf{x}, \mathbf{x})}.$$

Osoittaja on $x$- ja $y$-lukujen otoskovarianssi, ja nimittäjä on $x$-lukujen otosvarianssi. PNS-suoran yhtälö $y = a + bx$ voidaan esittää myös kaavalla

$$y - \bar{y} = b(x - \bar{x}), \tag{9.6}$$

mistä nähdään, että se kulkee parien $(x_i, y_i)$ keskiarvon $(\bar{x}, \bar{y})$ kautta. Tämä muotoilu on helpompi muistaa kuin PNS-suoran vakiotermin kaava (9.3), ja tästä muotoilusta nähdään helposti oikea lauseke PNS-suoran vakiotermille.

**Esimerkki 9.1** R:ssä vektorien `x` ja `y` otoskovarianssi saadaan laskettua kutsulla `cov(x, y)`, ja vektorin otoskovarianssi saadaan laskettua joko kutsulla `var(x)` (tai kutsulla `cov(x, x)`). Tällä keinolla PNS-suoran kertoimet saadaan laskettua itse helposti. Otoskovarianssin saa toki laskettua myös suoraan määritelmää käyttämällä.

```
x <- c(11, 12, 13, 13, 12, 11, 10.5, 10.5, 11.5, 11.5, 12.5,
    12.5)
y <- c(120, 153, 217, 177, 133, 90, 69, 55, 108, 144, 154, 179)
print(c(mean(x), mean(y), cov(x, y), var(x)))

## [1]  11.7500 133.2500  39.8409   0.7955

print(omaSxy <- 1/(length(x) - 1) * sum((x - mean(x)) * (y -
    mean(y))))

## [1] 39.84

print(b <- cov(x, y)/var(x))

## [1] 50.09

print(a <- mean(y) - b * mean(x))

## [1] -455.3
```

Toki R:stä löyty myös valmis funktio `lm`, jolla PNS-suora saadaan helposti laskettua.

```
print(lm(y ~ x))

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      -455.3         50.1
```

△

## PNS-suoran kertoimien kaavojen todistus

Todistamme kaavat (9.2) ja (9.3) suoralla laskulla. Todistus perustuu siihen tosiasiaan, että $\mathrm{SS}(\alpha, \beta)$ on kertoimien suhteen toisen asteen polynomi, jota voidaan analysoida neliöksi täydentämällä.

Lähdemme liikkeelle esityksestä

$$y_i - \alpha - \beta x_i = (y_i - \bar{y}) - (\alpha - \bar{y} + \beta\bar{x}) - \beta(x_i - \bar{x}),$$

jonka avulla saamme

$$\begin{aligned}\mathrm{SS}(\alpha,\beta) &= \sum_{i=1}^{n}\left[(y_i-\bar{y})-(\alpha-\bar{y}+\beta\bar{x})-\beta(x_i-\bar{x})\right]^2\\ &= \sum(y_i-\bar{y})^2 + n(\alpha-\bar{y}+\beta\bar{x})^2+\beta^2\sum(x_i-\bar{x})^2\\ &\quad -2(\alpha-\bar{y}+\beta\bar{x})\sum(y_i-\bar{y})\\ &\quad -2\beta\sum(x_i-\bar{x})(y_i-\bar{y})\\ &\quad +2(\alpha-\bar{y}+\beta\bar{x})\beta\sum(x_i-\bar{x}).\end{aligned}$$

Tässä kerrottiin trinomin neliö auki, ja summaoperaattorin alta otetiin pois vakiona pysyvät termit. Koska

$$\sum(x_i-\bar{x})=\sum x_i - n\bar{x}=0, \quad \text{ja} \quad \sum(y_i-\bar{y})=\sum y_i - n\bar{y}=0,$$

niin kaksi funktion $\mathrm{SS}(\alpha,\beta)$ esityksen kuudesta termistä häviää. Otamme käyttöön merkinnät

$$q_{xx}=\sum(x_i-\bar{x})^2, \quad q_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y}), \quad q_{yy}=\sum(y_i-\bar{y})^2, \tag{9.7}$$

minkä jälkeen $\mathrm{SS}(\alpha,\beta)$ voidaan esittää kaavalla

$$\begin{aligned}\mathrm{SS}(\alpha,\beta) &= q_{yy}+n(\alpha-\bar{y}+\beta\bar{x})^2\\ &\qquad q_{xx}\left(\beta^2-2\beta\frac{q_{xy}}{q_{xx}}+\frac{q_{xy}^2}{q_{xx}^2}\right)-\frac{q_{xy}^2}{q_{xx}}\\ &= q_{yy}-\frac{q_{xy}^2}{q_{xx}}+q_{xx}\left(\beta-\frac{q_{xy}}{q_{xx}}\right)^2+n(\alpha-\bar{y}+\beta\bar{x})^2.\end{aligned}$$

Koska oletusten mukaan $q_{xx}>0$ ja $n>0$, niin kaikilla $(\alpha,\beta)$ pätee

$$\mathrm{SS}(\alpha,\beta)\geq q_{yy}-\frac{q_{xy}^2}{q_{xx}}=\mathrm{SS}(a,b) \tag{9.8}$$

jossa alaraja saavutetaan valitsemalla $\beta$:lle kaavan (9.2) mukainen arvo $b$ ja sen jälkeen $\alpha$:lle kaavan (9.3) mukainen arvo $a$. Nyt on PNS-suoran kertoimien kaavat saatu todistettua.

## 9.3 Lineaarinen malli

Pienimmän neliösumman periaate voidaan johtaa suurimman uskottavuuden periaatteesta, kun ensin asetetaan lineaarinen malli, jossa virheet ovat normaalijakautuneita. Lineaarisessa mallissa havaintoja $y_1,\ldots,y_n$ vastaa satunnaismuuttujat $Y_1,\ldots,Y_n$, joiden yhteisjakauma voidaan esittää kaavalla

$$Y_i=\alpha+\beta\,x_i+\epsilon_i, \qquad i=1,\ldots,n. \tag{9.9}$$

Tässä ajatellaan, että luvut $x_1,\ldots,x_n$ ovat tunnettuja vakioita ja että virheet $\epsilon_i$ ovat riippumattomia satunnaismuuttujia, jotka kaikki noudattavat normaalijakaumaa

$$\epsilon_i\sim N(0,\sigma^2).$$

Virhesatunnaismuuttujien odotusarvo on nolla ja varianssi $\sigma^2$ ei riipu indeksistä $i$. Tuntemattomia parametreja ovat kertoimet $\alpha$ ja $\beta$ sekä virhevarianssi $\sigma^2 > 0$.

Lineaarinen malli (9.9) voidaan yhtäpitävästi muotoilla siten, että satunnaismuuttujat $Y_i$ ovat riippumattomia siten, että muuttujan $Y_i$ jakauma on

$$Y_i \sim N(\alpha + \beta\, x_i, \sigma^2), \qquad i = 1, \ldots, n. \tag{9.10}$$

Tämän huomion jälkeen uskottavuusfunktio voidaan kirjoittaa.

Kun $2\pi$:n potenssit jätetään pois, niin uskottavuusfunktioksi saadaan

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma^2}\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2\right). \end{aligned}$$

Logaritminen uskottavuusfunktio on

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 \tag{9.11}$$

$$= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \operatorname{SS}(\alpha, \beta), \tag{9.12}$$

jossa $\operatorname{SS}(\alpha, \beta)$ on virheiden neliösumma (9.1). Oli varianssiparametrin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion

$$(\alpha, \beta) \mapsto \ell(\alpha, \beta, \sigma^2)$$

maksimipiste on sama kuin funktion $\operatorname{SS}(\alpha, \beta)$ minimipiste, joka puolestaan on $(a, b)$, jossa $a$ ja $b$ ovat PNS-suoran vakiotermi ja kulmakerroin. Tämän ansiosta SU-estimaatin haku saadaan palautettua yhden muuttujan maksimointitehtäväksi, jossa pitää etsiä funktion

$$u(\sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \operatorname{SS}(a, b), \qquad \sigma^2 > 0$$

maksimipiste. Tämän funktion maksimi löytyy pisteestä

$$\hat{\sigma}^2 = \frac{1}{n} \operatorname{SS}(a, b).$$

Kertoimien SU-estimaateiksi saatiin PNS-suoran kertoimet, ja varianssiparametrin SU-estimaatiksi saatiin

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - a - b\, x_i)^2.$$

Lineaarisessa mallissa (9.9) kertoimien estimaatteina käytetään PNS-estimaatteeja $a$ ja $b$, mutta virhevarianssin $\sigma^2$ estimaattina ei käytetä SU-estimaattia, vaan estimaattia

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a - b\, x_i)^2, \tag{9.13}$$

jossa jakajana on otoskoon $n$ sekä mallin kertoimien lukumäärän 2 erotus.

Vastaavien estimaattorien $a(\mathbf{Y})$, $b(\mathbf{Y})$ ja $s^2(\mathbf{Y})$ otantajakauma tunnetaan. On suhteellisen yksinkertaista nähdä, että kertoimien PNS-estimaattorit ovat harhattomia ja normaalijakautuneita. Lyhyehkö lasku näyttää, että

$$b(\mathbf{Y}) \sim N(\beta, \frac{\sigma^2}{q_{xx}}). \tag{9.14}$$

Tämän takia PNS-kulmakertoimen $b$ keskivirhe lasketaan kaavalla

$$\mathrm{se}(b) = \frac{s}{\sqrt{q_{xx}}}, \tag{9.15}$$

jossa käytetään kaavan (9.7) merkintää neliösummalle

$$q_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

jota ei jaeta $(n-1)$:llä.

On paljon vaativampaa näyttää, että $s^2(\mathbf{Y})$ harhaton sekä riippumaton sekä estimaattorista $a(\mathbf{Y})$ että estimaattorista $b(\mathbf{Y})$. Osoittautuu, että sopivasti skaalattuna $s^2(\mathbf{Y})$ noudattaa khiin neliön jakaumaa vapausasteluvulla $n-2$, eli

$$\frac{n-2}{\sigma^2}\, s^2(\mathbf{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - a(\mathbf{Y}) - b(\mathbf{Y})\, x_i)^2 \sim \chi^2_{n-2}. \tag{9.16}$$

Kun otantajakaumat (9.14) ja (9.16) yhdistetään sen tiedon kanssa, että nämä estimaattorit ovat riippumattomia, niin nähdään että

$$\frac{b(\mathbf{Y}) - \beta}{s(\mathbf{Y})/\sqrt{q_{xx}}} \sim t_{n-2}. \tag{9.17}$$

Näihin tuloksiin perustuen voidaan mallin kulmakertoimelle laskea luottamusvälejä sekä voidaan testata sitä koskevia hypoteeseja.

Esimerkiksi $\beta$:n kaksisuuntainen luottamusväli luottamustasolla $(1-\alpha)$ lasketaan kaavalla

$$b - t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}} \leq \beta \leq b + t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}}. \tag{9.18}$$

On mielekästä testata hypoteesiparia

$$H_0 : \beta = 0, \qquad H_0 \neq 0,$$

sillä jos todellinen kulmakerroin on nolla, niin tällöin selittäjää $x$ ei lainkaan tarvita mallissa. Tässä testissä aineistosta lasketaan testisuure

$$t = \frac{b - 0}{s/\sqrt{q_{xx}}}, \tag{9.19}$$

jota verrataan $t_{n-2}$-jakauman yläkvantiiliin $t_{n-2}(\alpha/2)$. Nollahypoteesi hylätään, jos $|t| > t_{n-2}(\alpha/2)$.

Esimerkiksi kuminauha-aineistolle tämän testin tulos nähdään tutkimalla seuraavien komentojen tulostusta.

```
m <- lm(y ~ x)
summary(m)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.86 -13.49  -3.66  11.43  24.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -455.3       68.3   -6.66  5.6e-05 ***
## x               50.1        5.8    8.64  6.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.2 on 10 degrees of freedom
## Multiple R-squared: 0.882,Adjusted R-squared: 0.87
## F-statistic: 74.6 on 1 and 10 DF,  p-value: 5.98e-06
```

Nollahypoteesia $H_0 : \beta = 0$ vastaava $t$-arvo ja $p$-arvo ovat

$$t = 8.64, \qquad p = 5.98 \times 10^{-6}.$$

Tämä hypoteesi hylätään kaikilla tavanomaisilla luottamustasoilla. Toisin sanoen venytys on tarpeellinen selittäjä tässä lineaarisessa mallissa.

Kulmakertoimen 95 % luottamusvälin saa helpoimmin laskettua komennolla

```
confint(m)

##               2.5 %  97.5 %
## (Intercept) -607.47 -303.04
## x             37.17   63.01
```

Kulmakertoimen 95 % luottamusväli on $[37.17, 63.01]$.

## 9.4 Lineaarinen regressio, kun selittäjät ovat satunnaismuuttujia

Edellisen jakson tilastollinen malli on mielekäs lähinnä silloin, kun selittäjän arvot voidaan kokeessa itse asettaa, koska ne oletetaan tunnetuiksi vakioiksi. Lineaarista regressiota käytetään kuitenkin myös sellaisissa tilanteissa, joissa $(x, y)$-arvot mitataan yksilöistä, jotka on poimittu satunnaisesti populaatiosta. Tällöin myös $x$-arvot pitää mallintaa satunnaismuuttujina.

Jos tehdään mallissa sellainen oletus, että satunnaismuuttujat

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

muodostavat satunnaisotoksen jostakin kaksiulotteisesta normaalijakaumasta, niin tällöin edellisen jakson päättelyn kannalta keskeiset jakaumatulokset pitävät edelleen paikkansa. Erityisesti varianssiparametrin estimaattorin otantajakaumatulos (9.16) pitää edelleen paikkansa ja $t$-tunnusluvun (9.17) jakauma on edelleen $t_{n-2}$.

Tämä väite voidaan perustella tarkastelemalla sitä ehdollista jakaumaa, joka syntyy, kun selittäjävektorin $\mathbf{X} = (X_1, \ldots, X_n)$ arvoksi kiinnitetään havaitut arvot $\mathbf{x} = (x_1, \ldots, x_n)$. Satunnaismuuttujien $Y_1, \ldots, Y_n$ ehdollisessa yhteisjakaumassa ne ovat riippumattomia ja noudattavat normaalijakaumia

$$Y_i \mid (\mathbf{X} = \mathbf{x}) \sim N(\alpha + \beta\, x_i, \sigma^2),$$

jossa kertoimet $\alpha$ ja $\beta$ sekä virhevarianssi $\sigma^2$ voidaan ilmaista kaksiulotteisen normaalijakauman parametrien avulla. Tämä seuraa kaksiulotteisen normaalijakauman erityisominaisuuksista.

Siis ehdollinen yhteisjakauma täyttää lineaarisen mallin oletukset, joten kaikki edellisessä jaksossa mainitut jakaumatulokset pätevät ehdolla $\mathbf{X} = \mathbf{x}$. Tästä seuraa edelleen, että sellaiset jakaumatulokset, jotka eivät riipu arvosta $\mathbf{x}$ (erityisesti (9.16) sekä (9.17)) pätevät myös ei-ehdollisesti.

## 9.5 Muita lineaarisia malleja

Jakson 9.3 malli (9.9) voidaan kirjoittaa matriisimerkinnöillä kaavalla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{9.20}$$

Tässä asetelmamatriisi (tai mallimatriisi) $\mathbf{X}$ on kiinteä ja tunnettu, ja kerroinvektori $\boldsymbol{\beta}$ kiinteä mutta tuntematon. Mallille (9.9) ne ovat muotoa

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Virhevektorin $\boldsymbol{\epsilon}$ komponentit $\epsilon_i$ ovat riippumattomia ja noudattavat kukin normaalijakaumaa $N(0, \sigma^2)$. Tässä mallissa sekä kerroinvektori että virhevarianssi ovat tuntemattomia parametreja.

Usea tällä kurssilla käsitelty malli voidaan lineaarisena mallina kaavalla (9.20). Näitä ovat

- lineaarinen regressiomalli (9.9)
- satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, jossa $\mu$ ja $\sigma^2$ ovat tuntemattomia;
- kaksi riippumatonta otosta kahdesta eri normaalijakaumasta $N(\mu_1, \sigma^2)$ ja $N(\mu_2, \sigma^2)$, (varianssianalyysimalli), kun jakaumilla on sama tuntematon varianssiparametri.

Lineaaristen mallien teoria antaa yhtenäisen teorian kaikille näille malleille sekä kaikille niiden muotoa (9.20) oleville yleistyksille. Aiheeseen voi perehtyä esim. lineaaristen mallien kurssilla.

# Luku 10

# Bayes-päättelyn alkeita

Bayesiläisessä päättelyssä havaintosatunnaisvektorin jakauma mallinnetaan täysin samalla tavalla kuin frekventistisessä lähestymistavassa silloin, kun parametrin arvo on kiinnitetty. Frekventistisessä lähestymistavassa parametri on kiinteä (ts. ei-satunnainen) mutta tuntematon. Bayesiläisessä lähestymistavassa parametria käsitellään satunnaisena suureena.

Tämä ehkä vähäiseltä tuntuva ero johtaa suuriin eroihin laskutekniikoissa ja tulosten tulkinnoissa. *Bayesiläisessä päättelyssä kaikki on toisin* kuin frekventistisessä.

## 10.1 Todennäköisyyslaskentaa

Tämän jakson kaavoissa oletetaan hiljaisesti, että osamäärien nimittäjät ovat erisuuria kuin nolla.

Olkoot $A$ ja $B$ tapahtumia. Jos tiedämme (täsmälleen sen), että $B$ on sattunut, niin tapahtuman $A$ todennäköisyys lasketaan *ehdollisen todennäköisyyden* kaavalla

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{10.1}$$

Ehdollisen todennäköisyyden kaavasta saadaan todennäköisyyksien *kertolaskukaava*

$$P(A \cap B) = P(B)\, P(A \mid B) = P(A)\, P(B \mid A). \tag{10.2}$$

Tästä nähdään kaava

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}. \tag{10.3}$$

Jos tapahtumat $A_1, \ldots, A_M$ ovat jokin perusjoukon ositus ts.

- joukot $A_i$ ovat erillisiä: jos $i \neq j$, niin $A_i \cap A_j = \emptyset$.
- niiden yhdiste on koko perusjoukko,

niin $P(B)$ voidaan laskea (todennäköisyyden additiivisuuden perusteella) seuraavasti,

$$P(B) = P(\cup_{i=1}^{M}(B \cap A_i)) = \sum_{i=1}^{M} P(B \cap A_i)$$
$$= \sum_{i=1}^{M} P(A_i)\, P(B \mid A_i)$$

Kun tätä ns. *kokonaistodennäköisyyden* kaavaa käytetään kaavassa (10.3) saadaan *Bayesin kaava*

$$P(A_k \mid B) = \frac{P(A_k)\, P(B \mid A_k)}{\sum_{i=1}^{M} P(A_i)\, P(B \mid A_i)} \tag{10.4}$$

johon bayesiläinen päättely perustuu diskreetin parametrin ja diskreetin havaintovektorin tapauksessa.

Kirjoitetaan edelliset kaavat vielä siinä tapauksessa, jossa käsitellään kahden diskreetin satunnaismuuttujan (tai satunnaisvektorin) $\tilde{\theta}$ ja $\mathbf{Y}$ yhteisjakaumaa, jonka määrää niiden yhteispistetodennäköisyysfunktio

$$f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}) = P(\tilde{\theta} = \theta, \mathbf{Y} = \mathbf{y}) = P(\{\tilde{\theta} = \theta\} \cap \{\mathbf{Y} = \mathbf{y}\}). \tag{10.5}$$

Satunnaismuuttujan (tai satunnaisvektorin) $\tilde{\theta}$ mahdolliset arvot ovat $\theta_1, \theta_2, \ldots$ ja satunnaismuuttujan (tai satunnaisvektorin) $\mathbf{Y}$ mahdolliset arvot ovat $\mathbf{y}_1, \mathbf{y}_2, \ldots$. Yhteispistetodennäköisyysfunktion arvoja ja muiden pistetodennäköisyysfunktioiden arvoja lasketaan jatkossa sellaisissa pisteissä $(\theta, \mathbf{y})$, joiden $\theta$-koordinaatti on jokin $\tilde{\theta}$ mahdollisista arvoista ja $\mathbf{y}$-koordinaatti on jokin $\mathbf{Y}$:n mahdollisista arvoista.

Käytämme seuraavia merkintöjä.

- $p(\theta)$ on satunnaismuuttujan $\tilde{\theta}$ reunajakauman ptnf, eli
$$p(\theta) = P(\tilde{\theta} = \theta)$$
- $f(\mathbf{y} \mid \theta)$ on satunnaisvektorin $\mathbf{Y}$ pntf, kun $\tilde{\theta} = \theta$, eli
$$f(\mathbf{y} \mid \theta) = P(\mathbf{Y} = \mathbf{y} \mid \tilde{\theta} = \theta).$$
- $p(\theta \mid \mathbf{y})$ on satunnaismuuttujan $\tilde{\theta}$ ptnf, kun $\mathbf{Y} = \mathbf{y}$, eli
$$p(\theta \mid \mathbf{y}) = P(\tilde{\theta} = \theta \mid \mathbf{Y} = \mathbf{y})$$
- $f(\mathbf{y})$ on satunnaisvektorin $\mathbf{Y}$ reunajakauman ptnf, eli
$$f(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y})$$

Satunnaismuuttujien $\tilde{\theta}$ ja $\mathbf{Y}$ reunapistetodennäköisyysfunktiot saadaan niiden yhteispistetodennäköisyysfunktiosta kokonaistodennäköisyyden kaavalla, nimittäin

$$p(\theta) = P(\tilde{\theta} = \theta) = \sum_{\mathbf{y}} P(\tilde{\theta} = \theta, \mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{y}} f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}) \tag{10.6}$$

$$f(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}) = \sum_{\theta} f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}). \tag{10.7}$$

Todennäköisyyksien kertolaskukaavan (10.2) mukaan yhteispistetodennäköisyysfunktio voidaan jakaa tekijöihin molemmilla seuraavista tavoista

$$f_{\tilde{\theta},\mathbf{Y}}(\theta,\mathbf{y}) = p(\theta)\, f(\mathbf{y} \mid \theta) = f(\mathbf{y})\, p(\theta \mid \mathbf{y}). \tag{10.8}$$

Tapahtuman $\tilde{\theta} = \theta$ todennäköisyys ehdolla $\mathbf{Y} = \mathbf{y}$ saadaan ratkaistua edellisestä identiteetistä, nimittäin

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})}. \tag{10.9}$$

Tämä on se Bayesin kaavan versio, johon bayesiläinen päättely perustuu silloin kuin aineiston otantajakauma on diskreetti ja parametriavaruus on diskreetti. Tässä parametria pidetään satunnaismuuttujana $\tilde{\theta}$ joka saa jonkin arvon parametriavaruudessa $\Theta$.

Ennen (lat. *a priori*) havaintojen tekoa parametrilla on ns. *priorijakauma* (engl. *prior distribution*) eli jakauma, jonka ptnf on $p(\theta)$. Seuraavaksi tehdään havainto $\mathbf{Y} = \mathbf{y}$. Bayesiläinen päättely tarkoittaa sitä, että havaintojen jälkeen (lat. *a posteriori*) priorikäsitys päivitetään siirtymällä havaintoja vastaavaan ehdolliseen jakaumaan. Ennakkokäsitys päivitetään Bayesin kaavalla havaintojen jälkeiseksi käsitykseksi käyttämällä hyväksi priorijakaumaa ja havaintoja vastaavaa uskottavuusfunktiota $f(\mathbf{y} \mid \theta)$. Tulos on parametrin *posteriorijakauma* (engl. *posterior distribution*), eli parametrin ehdollinen jakauma $p(\theta \mid \mathbf{y})$ ehdolla $\mathbf{Y} = \mathbf{y}$.

Bayesin kaava kannattaa pitää mielessä muodossa

$$\text{posteriori} \propto \text{priori} \times \text{uskottavuus} \tag{10.10}$$

Tässä merkintä $\propto$ tarkoittaa verrannollisuutta, ts. edellä väitetään, että posteriori on vakio kertaa priorin ja uskottavuusfunktion tulo. Tässä posterioria $p(\theta \mid \mathbf{y})$ ajatellaan muuttujan $\theta$ funktiona kuten myös priorin ja uskottavuusfunktion tuloa. Ts. edellä väitetään, että

$$p(\theta \mid \mathbf{y}) = C\, p(\theta)\, f(\mathbf{y} \mid \theta), \qquad \text{kaikilla } \theta, \tag{10.11}$$

ja tämän on totta, sillä

$$C = \frac{1}{f(\mathbf{y})} = \frac{1}{\sum_{\theta=0}^{N} p(\theta)\, f(\mathbf{y} \mid \theta)}$$

Verrannollisuusvakio $C$ toki riippuu havainnoista $\mathbf{y}$, mutta muuttujan $\theta$ funktiona ajateltuna se on vakio.

Bayesiläisessä analyysissä havaintoja vastaavalle satunnaisvektorille kiinnitetään sen havaittu arvo, ja sitten pohditaan eri parametrinarvojen todennäköisyyksiä. Frekventistisessä analyysissä parametri on kiinteä, ja todennäköisyyslaskentaa käytetään aineistoa vastaavan satunnaisvektorin jakauman ja siitä johdettujen tunnuslukujen jakaumien johtamiseen erilaisilla hypoteettisilla parametrinarvoilla. Useimmat frekventistisen tilastotieteen käsitteet perustuvat sellaisten aineistojen ominaisuuksien pohtimiseen, joita ei kokeessa havaittu.

Voimme ajatella, että bayesiläisessä analyysissä diskreetin parametrin tapauksessa lasketaan priorin ja uskottavuuden tulo kaikilla mahdollisilla parametrin arvoilla, jonka jälkeen tulos normalisoidaan pistetodennäköisyysfunktioksi jakamalla laskettujen arvojen summalla. Tämän algoritmin vaiheet ovat

1. Laske
$$s = \sum_{\theta} p(\theta)\, f(\mathbf{y} \mid \theta). \tag{10.12}$$
2. Laske
$$p(\theta \mid \mathbf{y}) = \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{s}, \qquad \theta \in \Theta. \tag{10.13}$$

Näemme sovelluksia seuraavissa jaksossa, joissa palaamme käsittelemään tilastollista päättelyä luvun 2 esimerkeissä.

## 10.2 Pallot kulhossa: diskreetti parametri

Kulhossa on $N$ palloa: $\theta$ valkoista palloa ja $N - \theta$ mustaa. Valkoisten pallojen lukumäärä $0 \leq \theta \leq N$ on tuntematon. Palloja nostetaan satunnaisesti ja palauttaen $n$ kertaa.

Oletetaan nyt, että luontoäiti on laittanut pallot kulhoon sillä tavalla, että hän ensin arpoi valkoisten pallojen lukumääräksi $\theta$ yhden luvuista $0, 1, \ldots, N$ siten, että kaikki vaihtoehdot ovat yhtä todennäköisiä. Sen jälkeen hän laittoi kulhoon $\theta$ valkoista ja $N - \theta$ mustaa palloja.

Pallojen poiminta tuottaa jonkin jonon $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ onnistumisia (valkoisen pallon nosto koodataan arvolla $y_i = 1$) tai epäonnistumisia (mustan pallon nosto koodataan arvolla $y_i = 0$). Vastaavien satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteispistetodennäköisyysfunktio tunnetaan, jos $\theta$ tunnetaan, sillä se on
$$f(\mathbf{y} \mid \theta) = \left(\frac{\theta}{N}\right)^{t(\mathbf{y})} \left(1 - \frac{\theta}{N}\right)^{n - t(\mathbf{y})}, \tag{10.14}$$
jossa $t(\mathbf{y}) = \sum_i y_i$ on onnistumisten lukumäärä.

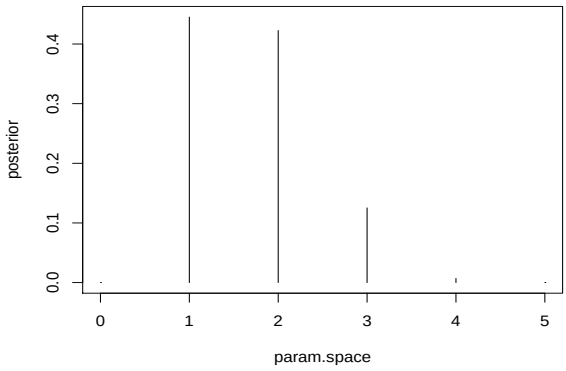
Kun on havaittu tietty jono $\mathbf{y}$, niin sitten voidaan kysyä, millä todennäköisyydellä kulhossa on $\theta$ kappaletta valkoisia palloja. Kun tähän kysymykseen vastataan kaikilla mahdollisilla parametrin $\theta$ arvoilla $0, 1, \ldots, N$, saadaan tuloksena posteriorijakauma.

Ennen havaintoja satunnaismuuttujan $\tilde{\theta}$ jakauma eli priorijakauma on tilanteen kuvauksen perusteella diskreetti tasajakauma, jonka pistetodennäköisyysfunktio on
$$p(\theta) = \frac{1}{N+1}, \qquad \theta = 0, 1, \ldots, N.$$
Posteriorijakauma lasketaan seuraavassa esimerkissä soveltamalla kaavoja (10.12)–(10.13).

**Esimerkki 10.1** Kulhossa on $N = 5$ palloa ja nostoja tehdään $n = 7$ ja tulokset ovat $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$, jolloin onnistumisia on $x = 2$ kappaletta. Lasketaan R:llä priorin ja uskottavuusfunktion tulo, ja normalisoidaan se posteriorijakaumaksi.

```
N <- 5
n <- 7
x <- 2
param.space <- 0:N
print(prior <- rep(1, N + 1)/(N + 1))
```

```
## [1] 0.1667 0.1667 0.1667 0.1667 0.1667 0.1667

print(likelihood <- (param.space/N)^x * (1 - param.space/N)^(n -
    x))

## [1] 0.0000000 0.0131072 0.0124416 0.0036864 0.0002048 0.0000000

h <- prior * likelihood
posterior <- h/sum(h)
names(posterior) <- as.character(param.space)
print(posterior)

##        0        1        2        3        4        5
## 0.000000 0.445217 0.422609 0.125217 0.006957 0.000000

plot(param.space, posterior, "h")
```




Tarkkaavainen lukija huomasi, että tasaisesti priorista seurasi se, että posteriorijakauma oli yhtä kuin uskottavuusfunktio normalisoituna todennäköisyysjakaumaksi. Havaintojen jälkeen parametrin todennäköisin arvo on 1, mutta arvo 2 on lähes yhtä todennäköinen. Tässä tilanteessa on on paikallaan puhua parametrin todennäköisyydestä (ennen ja jälkeen havaintojen teon); tässä yhteydessä ei tarvitse puhua esim. uskottavuudesta. △

## 10.3 Priorin ja posteriorin tulkitseminen epävarmuuden kuvauksina

Edellisen jakson laskuissa ei ole mitään kiistanalaista, vaan ne seuraavat suoraan todennäköisyyslaskennan sääntöjen avulla tilanteen kuvauksesta. Kuitenkin bayesiläistä päättelyä pidettiin tilastotieteilijöiden piirissä yleisesti ainakin 1920–1960 luvuilla vanhentuneena ja suorastaan tuomittavana tapana lähestyä tilastollisen päättelyn ongelmia. Nykypäivänä tilastotieteen ammattilehdissä suurin osa artikkeleistä käyttää tavalla tai toisella bayesiläistä lähestymistapaa.

Yritän tässä jaksossa valottaa, mikä asia bayesiläisessä päättelyssä aikanaan aiheutti tämän voimakkaan vastustuksen.

Ongelmaksi koettiin se, että bayesiläistä lähestymistapaa sovellettiin tilanteissa, joissa parametrin arvoa ei määrännyt mikään satunnaismekanismi. Tällöin priorijakauma ei kuvaa todellista arpomista, vaan se pitää ymmärtää kvantitatiivisena kuvauksena soveltajan epävarmuudesta parametrin todellisesta arvosta ennen havaintojen tekemistä. Posteriorijakauman tulkinnaksi tulee puolestaan se, että se on kvantitatiivinen kuvaus menetelmän soveltajan epävarmuudesta parametrin todellisesta arvosta, kun havainnot on otettu huomioon. Kiistan ydin oli siinä, että frekventistisen tilastotieteen perustajat eivät hyväksyneet sitä ajatusta, että todennäköisyysjakauma saataisiin tulkita subjektiivisena epävarmuuden kuvauksena. Heidän mielestään todennäköisyyden käsite oli objektiivinen, ja sitä saadaan käyttää ainoastaan sellaisissa tilanteessa, jossa jotakin koetta (jossakin mielessä) toistetaan useita kertoja.

Nykyaikana bayesiläisessä päättelyssä lähes aina käytetään todennäköisyyden subjektiivista tulkintaa epävarmuuden kvantitatiivisena kuvauksena silloin, kun puhutaan parametrin todennäköisyysjakaumasta. Tätä ajatusta ei nykyään enää koeta ongelmallisena.

Posteriorijakauma on tällöin tietenkin subjektiivinen, sillä se riippuu siitä, minkälaista priorijakaumaa kyseinen subjekti pitää hyvänä kuvauksena omasta epävarmuudestaan. Priorijakaumalla on voimakas vaikutus posteriorijakaumaan, mikäli otoskoko on pieni. Jos otoskoko on suuri, niin tällöin erilaisilla järkevillä prioreilla saavutetaan lähes samanlainen posteriorijakauma. Otoskoon kasvaessa järkevien soveltajien subjektiiviset posteriorijakaumat alkavat siis muistuttaa yhä enenenevässä määrin toisiaan.

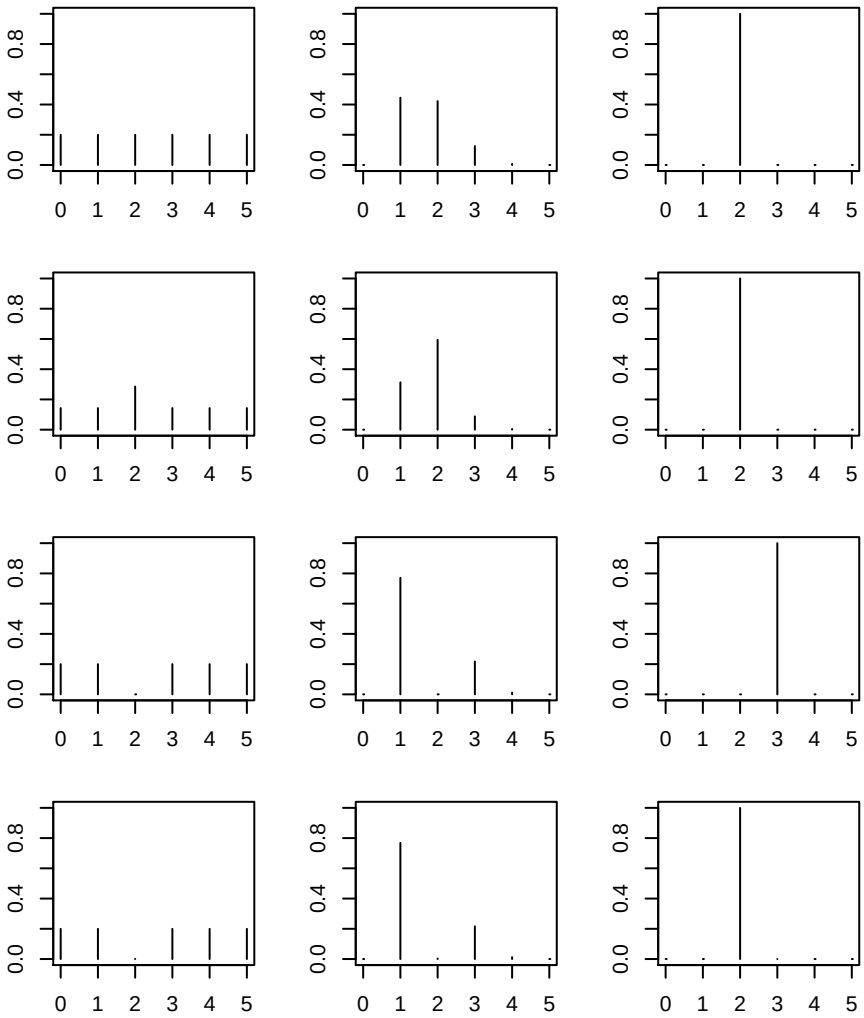
Lasketaan esimerkin vuoksi neljän eri henkilön A, B, C ja D posteriorijakaumat pallot kulhossa -tilanteessa: ensin seitsemän noston (2 onnistumista) ja sitten 300 noston (133 onnistumista) jälkeen.

- A:n priorijakauma on tasajakauma arvoilla $0, 1, \ldots, 5$.
- B suosii sitä vaihtoehto, että kulhossa on kaksi valkoista palloa. B:n priorijakauman mukaan valkoisten pallojen lukumäärä 2 on kaksi kertaa todennäköisempi kuin muut, jotka puolestaan ovat keskenään yhtä todennäköisiä.
- C on dogmaattisesti sitä mieltä, että kulhossa ei voi olla kahta valkoista palloa. C:n ennakkokäsityksen mukaan arvo 2 on mahdoton, ja kaikki muut mahdollisuudet ovat yhtä todennäköisiä.
- Myös D vastustaa kiivaasti sitä ajatusta, että kulhossa voisi olla kaksi valkoista palloa. Hän kuitenkin muotoilee käsityksensä vähemmän dogmaattisesti kuin C. D:n ennakkokäsityksen mukaan arvolla 2 on todennäköisyys 1/1000, ja kaikki muut arvot ovat keskenään yhtä todennäköisiä.

Eri henkilöiden priorijakaumat ja posteriorijakaumat on esitetty kuvassa 10.3

Otoskoolla $n = 7$ nähdään, että posteriorijakaumat riippuvat vahvasti kunkin henkilön priorikäsityksistä, mutta henkilöt C ja D ovat käytännössä yhtä mieltä eri valkoisten pallojen lukumäärän todennäköisyyksistä. Sen sijaan otoskoolla $n = 300$ kaikkien muiden henkilöiden paitsi C:n posteriorikäsitykset ovat käytännössä yhtenevät. C sulkee omalla priorin valinnallaan kokonaan pois sen mahdollisuuden, että valkoisia palloja voisi kulhossa olla kaksi. Tämän takia

**Kuva 10.1** Vaakariveillä on henkilöiden A, B, C ja D priorijakaumat, posteriorijakaumat $n = 7$ toiston ja 2 onnistumisen sekä $n = 300$ toiston ja 133 onnistumisen jälkeen.



C:n posterioritodennäköisyys kahdelle valkoiselle pallolle on aina nolla riippumatta lainkaan siitä, mitä havaintoja saadaan. Tällä kertaa aineisto todistaa otoskoolla $n = 300$ vahvasti sen puolesta, että valkoisia palloja olisi kulhossa todellisuudessa kaksi kappaletta, ja henkilöt A, B ja D ovat käytännössä täysin vakuuttuneita siitä, että valkoisia palloja on kulhossa kaksi.

Sellaista dogmaattista priorin valintaa (kuten C:n priori), joka sulkee kokonaan pois tarkastelusta jonkin järkevän osan parametriavaruudesta, voidaan pitää bayesiläisen päättelyn pelisääntöjen vastaisena.

## 10.4 Nasta purkissa: jatkuva parametri

Nyt binomikokeen onnistumistodennäköisyydellä on jokin arvo avoimella välillä $(0, 1)$. Jos tämä arvo $\theta$ tunnetaan, niin havaintosatunnaisvektorin pistetodennäköisyysfunktio on

$$f(\mathbf{y} \mid \theta) = \theta^{t(\mathbf{y})} \, (1 - \theta)^{n - t(\mathbf{y})},$$

kun $\mathbf{y} = (y_1, \ldots, y_n)$ on jono nollia tai ykkösiä, ja $t(\mathbf{y}) = \sum_i y_i$ on onnistumisten lukumäärä $n$ toistossa. Nyt parametriavaruus on jatkuva, ja tämä tuo mukanaan uusia teknisiä ongelmia.

Parametri $\tilde{\theta}$ on nyt jatkuvasti jakautunut satunnaismuuttuja, jonka arvot kuuluvat parametriavaruuteen $\Theta = (0, 1)$. Ennen havaintojen tekoa soveltajan pitää onnistua kuvaamaan oma epävarmuutensa parametrin arvoista priorijakaumalla, jonka tiheysfunktiota merkitsemme

$$p(\theta), \qquad \theta \in \Theta.$$

Havaintojen jälkeen siirrytään tarkastelemaan parametrin ehdollista tiheysfunktiota eli sen posteriorijakaumaa.

Suuri osa jakson 10.1 kaavoista pitää sellaisenaan paikkansa myös tässä uudessa tilanteessa, mutta summaus pitää korvata integroinnilla. Yhteisjakauma voidaan esittää funktion

$$f_{\tilde{\theta}, \mathbf{Y}}(\theta, \mathbf{y}) = p(\theta) \, f(\mathbf{y} \mid \theta)$$

avulla. Se on muuttujan $\theta$ suhteen tiheysfunktio ja muuttujan $\mathbf{y}$ suhteen pistetodennäköisyysfunktio, ts. todennäköisyyksiä lasketaan integroimalla muuttujan $\theta$ suhteen ja summaamalla muuttujan $\mathbf{y}$ suhteen.

Bayesin kaava saa tutun muodon

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta) \, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})},$$

mutta nyt normalisointivakio $f(\mathbf{y})$ eli havaintosatunnaisvektorin reunajakauman pistetodennäköisyysfunktion arvo pitää laskea integroimalla,

$$f(\mathbf{y}) = \int_0^1 p(\theta) \, f(\mathbf{y} \mid \theta) \, \mathrm{d}\theta.$$

Tätä integraalia ei yleisesti ottaen osata laskea analyyttisesti. Sitä voi toki yrittää approksimoida jollakin numeerisella menetelmällä.

Posteriorijakauman tiheysfunktion kuva voidaan piirtää suoraan käyttämällä verrannollisuustulosta

$$p(\theta \mid \mathbf{y}) \propto p(\theta) \, f(\mathbf{y} \mid \theta),$$

mikäli tyydytään siihen, että $y$-akselin skaala jää selvittämättä.

## 10.5 Liittojakauma eli konjugaattijakauma

Joillekin uskottavuusfunktioille on mahdollista löytää sellainen parametrinen perhe jakaumia, joilla on seuraava miellyttävä ominaisuus. Mikäli priorijakauma valitaan kyseisestä perheestä, niin myös posteriorijakauma kuuluu samaan perheeseen. Näemme kohta, että binomikokeen uskottavuusfunktiolle beeta-jakaumat muodostavat tällaisen perheen. Tämä voidaan ilmaista sanomalla, että beeta-jakauma on binomiuskottavuuden *liittopriori* (engl. *conjugate prior*) tai että havaintosatunnaisvektorin otantajakauma ja priorijakauma ovat toistensa liittojakaumia.

Mikäli tutkijan ennakkokäsitys parametrinarvosta voidaan esittää jollakin liittoperheen jakaumalla, niin tällöin posteriorijakauma saadaan laskettua johtamalla päivityskaavat, joilla priorijakauman parametrit päivitetään posteriorijakauman parametreiksi. Kutsutaan näitä liittojakaumaperheen parametreja selvyyden vuoksi hyperparametreiksi, jotta ne saadaan erotettua tilastollisen mallin parametrista $\theta$.

Beeta-jakauma on jatkuva jakauma, jonka tiheysfunktio on muotoa

$$g(x) \propto x^{\alpha-1}\,(1-x)^{\beta-1}, \quad \text{kun } 0 < x < 1, \tag{10.15}$$

missä $\alpha > 0$ ja $\beta > 0$ ovat jakaumaperheen parametrit. Tämä lauseke määrittelee yksikäsiteisesti tietyn todennäköisyysjakauman, jonka tiheysfunktio saadaan selville jakamalla lauseke sen välin $(0,1)$ yli lasketulla integraalilla, sillä tiheysfunktion integraalin koko satunnaismuuttujan arvoalueen yli täytyy olla yksi. Kaavassa (10.15) merkitsemättä jätetty normalisointivakio saadaan ns. Eulerin beetafunktion

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}\,(1-x)^{\beta-1}\,\mathrm{d}x$$

avulla (jossa $B$ on iso beeta-kirjain). Beeta-jakauman tiheysfunktion täydellinen kaava on

$$g(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\,x^{\alpha-1}\,(1-x)^{\beta-1}, & \text{kun } 0 < x < 1, \\ 0 & \text{muuten.} \end{cases} \tag{10.16}$$

Erityisesti valinnoilla $\alpha = 1$ ja $\beta = 1$ saadaan välin $(0,1)$ tasajakauma.

Beeta-jakauman $\mathrm{Beta}(\alpha,\beta)$ ominaisuudet tunnetaan. Sitä noudattavan satunnaismuuttujan $X$ arvot ovat välillä $(0,1)$, ja esim. sen odotusarvo ja varianssi saadaan laskettua tunnetuilla kaavoilla

$$EX = \frac{\alpha}{\alpha+\beta}, \qquad \operatorname{var} X = \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}. \tag{10.17}$$

Jakauman moodi (eli tiheysfunktion maksimipiste) on

$$\frac{\alpha-1}{\alpha+\beta-2},$$

mikäli $\alpha > 1$ ja $\beta > 1$ (muilla parametrinarvoilla jakaumalla ei ole hyvin määriteltyä moodia).

Tarkistetaan nyt, että beeta-jakauma on binomiuskottavuuden liittopriori. Priorin hyperparametrit ovat $\alpha, \beta > 0$ ja onnistumisia havaitaan $k$ kappaletta. Tarkistuksen voi tehdä kahdella erilaisella tavalla.

### Tapa 1: normalisointivakion laskeminen integroimalla

Integroidaan tulo priori kertaa uskottavuus. Huomaa, että priorijakauman tiheysfunktiossa argumenttina pitää käyttää integrointimuuttujaa, joka on $\theta$.

$$\begin{aligned} f(\mathbf{y}) &= \int_0^1 p(\theta)\, f(\mathbf{y} \mid \theta)\, \mathrm{d}\theta \\ &= \int_0^1 \frac{1}{B(\alpha,\beta)}\, \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\theta^{k}\,(1-\theta)^{n-k}\,\mathrm{d}\theta \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{\alpha+k-1}\,(1-\theta)^{\beta+n-k-1}\,\mathrm{d}\theta = \frac{B(\alpha+k,\beta+n-k)}{B(\alpha,\beta)}. \end{aligned}$$

Integraali osattiin laskea, koska huomattiin käyttää Eulerin beeta-funktion määritelmää.

Bayesin kaavan mukaan

$$\begin{aligned} p(\theta \mid \mathbf{y}) &= \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})} \\ &= \frac{\frac{1}{B(\alpha,\beta)}\, \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\theta^{k}\,(1-\theta)^{n-k}}{B(\alpha+k,\beta+n-k)/B(\alpha,\beta)} \\ &= \frac{1}{B(\alpha+k,\beta+n-k)}\, \theta^{\alpha+k-1}\,(1-\theta)^{\beta+n-k-1}, \qquad 0<\theta<1 \end{aligned}$$

Tästä nähdään, että posteriorijakauma on beeta-jakauma

$$\mathrm{Beta}(\alpha+k,\beta+n-k).$$

### Tapa 2: verrannollisuustarkastelu

Edellisessä tavassa tulee matkan varrella helposti virheitä. Tulos on paljon helpompi johtaa seuraavalla tekniikalla. Muuttujan $\theta$ funktiona posteriorijakauman tiheys on verrannollinen priorijakauman tiheyden ja uskottavuusfunktion tuloon, joten välillä $0<\theta<1$ pätee verrannollisuus

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto p(\theta)\, f(\mathbf{y} \mid \theta) \\ &= \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\theta^{k}\,(1-\theta)^{n-k} \\ &= \theta^{\alpha+k-1}\,(1-\theta)^{\beta+n-k-1}. \end{aligned}$$

Vertaamalla tulosta kaavaan (10.15) nähdään, että posteriorijakauma on beeta-jakauma

$$\mathrm{Beta}(\alpha+k,\beta+n-k),$$

sillä ainoa todennäköisyysjakauma, jonka kantaja on $(0,1)$ ja jonka tiheysfunktio on verrannollinen johdettuun funktioon on tämä beeta-jakauma.

## 10.6 Posteriorijakauman yhteenvetoja

Kun posteriorijakauma on selvitetty, niin lopuksi voidaan yrittää laskea siitä tiettyjä yhteenvetoja. Jos posteriorijakauma on yleisesti tunnettu yksinkertainen jakauma, niin tämä vaihe voidaan sivuuttaa.

Posteriorijakauman keskikohtaa voidaan luonnehtia parametrin posteriori-odotusarvolla. Binomikokeen tapauksessa tämä on luku

$$E[\tilde{\theta} \mid \mathbf{y}] = \int_0^1 \theta\, p(\theta \mid \mathbf{y})\, \mathrm{d}\theta.$$

Jos priori on Beta$(\alpha, \beta)$, niin posterioriodotusarvo voidaan lukea suoraan beetajakauman odotusarvon kaavasta (10.17), kun siihen sijoitetaan posteriorijakauman hyperparametrit, jotka ovat binomikokeessa

$$\alpha_1 = \alpha + k, \qquad \beta_1 = \beta + n - k.$$

Vastaavasti voidaan laskea posteriorimoodi eli posteriorijakauman moodi. Posterioriodotusarvoa ja posteriorimoodia voidaan ajatella bayesiläisinä piste-estimaatteina. Posteriorijakauman keskittyneisyyttä voidaan kuvailla esim. laskemalla parametrin posteriorivarianssi eli posteriorijakauman varianssi.

Mikäli posteriorijakauman kvantiilifunktion $q(u)$ arvoja osataan laskea tavalla tai toisella, niin sen jälkeen parametriavaruudesta löytyy helposti väli $[L, U]$, jolle

$$P(L \leq \tilde{\theta} \leq U \mid \mathbf{Y} = \mathbf{y}) = 1 - \alpha$$

millä tahansa annetulla $0 < \alpha < 1$. Eräs tällainen väli saadaan jakamalla virhetodennäköisyys $\alpha$ tasan alemman ja ylemmän jakauman hännän kesken valitsemalla

$$L = q(\alpha/2), \qquad U = q(1 - \alpha/2).$$

Havainnon jälkeen parametri kuuluu tälle välille todennäköisyydellä $1 - \alpha$. Tälläistä väliä voidaan kutsua tason $1 - \alpha$ todennäköisyysväliksi tai bayesiläiseksi luottamusväliksi.

Jos priorijakauma on beetajakauma, niin binomikokeessa posteriorijakauma on eräs toinen beetajakauma. Tilastollisista ohjelmistoista löytyy välineet beetajakauman kvantiilifunktion laskemiseksi: esim. R-ohjelmistossa kvantiilifunktion saa laskettua funktiolla `qbeta`. Tällaisessa tilanteessa todennäköisyysvälin päätepisteet saa laskettua käden käänteessä tietokoneella.

Mikäli ollaan kiinnostuneita muotoa

$$\tilde{\theta} \in A$$

olevasta hypoteesista, niin sen todennäköisyys havainnon jälkeen voidaan laskea integroimalla posteriorijakauman tiheysfunktiota

$$P(\tilde{\theta} \in A \mid \mathbf{Y} = \mathbf{y}) = \int_A p(\theta \mid \mathbf{y})\, \mathrm{d}\theta.$$

Jos posteriorijakauman kertymäfunktio osataan laskea ja jos $A$ on väli, niin välin $A$ posterioritodennäköisyys saadaan laskemalla kerymäfunktion arvot välin päätepisteissä sekä laskemalla suuremman arvon ja pienemmän arvon erotus. Vastahypoteesin $\tilde{\theta} \notin A$ todennäköisyys havainnon jälkeen saadaan sitten vähentämällä luvusta yksi tarkasteltavan hypoteesin todennäköisyys.

Jos priori on beetajakauma, niin posteriori on myös beetajakauma, ja tilastoohjelmista löytyy valmiudet laskea beetajakauman kertymäfunktion arvoja. Tässä tilanteessa välien posterioritodennäköisyydet saadaan laskettua käden käänteessä.

Koska parametria käsitellään satunnaismuuttujana, niin päästään puhumaan parametrin todennäköisyyksistä tai parametrin todennäköisyysjakaumasta tai kyseisen jakauman ominaisuuksista havainnon tekemisen jälkeen. Johdot ovat periaatteessa täysin suoraviivaisia ja niissä tarvitaan ainoastaan todennäköisyyslaskentaa, eikä mitään sen ulkopuolisia periaatteita. Monimutkaisemmille tilastollisille malleille vaadittavat laskut ovat kuitenkin liian hankalia analyyttisesti suoritettaviksi.

## 10.7 Bayesiläisen päättelyn laskentamenetelmiä

Posteriorijakauma saadaan selvitettyä kaavojen avulla seuraavissa tilanteissa.

1. Jos parametri on diskreetti ja sillä on äärellinen määrä mahdollisia arvoja.
2. Jos käytetään priorijakaumaa, joka kuuluu uskottavuusfunktion liittoperheeseen.

Muissa tapauksissa ei saada johdettua käyttökelpoisia analyyttisiä kaavoja.

Nykyään bayesiläinen analyysi tehdään tietokoneen avulla. Laskentamenetelmät perustuvat tyypillisesti satunnaisuuden simulointiin tietokoneen avulla eli ns. Monte Carlo -menetelmiin.

Perustana on se ajatus, että koska posteriorijakauma on todennäköisyysjakauma, niin sitä voidaan yrittää simuloida tietokoneella. On kehitetty menetelmiä, joissa lasketaan suuri määrä arvoja $t_1, \ldots, t_N$, joita voidaan pitää sellaisten satunnaismuuttujien $\tilde{\theta}_1, \ldots, \tilde{\theta}_N$ havaittuina arvoina, joista kukin on jakautunut posteriorijakauman mukaisesti.

Tämän jälkeen otosta $t_1, \ldots, t_N$ käsitellään data-analyysin keinoin.

- Posterioriodotusarvoa voidaan arvioida vastaavilla otoskeskiarvoilla,

$$E[\tilde{\theta} \mid \mathbf{y}] \approx \frac{1}{N} \sum_{i=1}^{N} t_i$$

- Parametrille voidaan muodostaa bayesiläisiä luottamusvälejä otoksesta laskettujen kvantiilipisteiden avulla.
- Posteriorijakauman pistetodennäköisyysfunktiota voidaan arvioida vastaavilla suhteellisilla frekvensseillä (diskreetin parametrin tapauksessa).
- Posteriorijakauman tiheysfunktiota voidaan arvioida esim. histogrammilla (jatkuvan parametrin tapauksessa).
- Mielivaltaisen parametrin funktion $g(\tilde{\theta})$ posteriorijakauman ominaisuuksia (esim. posterioriodotusarvoa tai posteriorijakaumaa) voidaan selvittää otoksesta $g(t_1), \ldots, g(t_N)$ aivan samoilla menetelmillä.

Tällaiset menetelmät ovat suhteellisen uusia: intensiivinen kehitysvaihe alkoi 1980-luvun lopulla. Niiden ansiosta nykyään on mahdollista analysoida lähes mielivaltaisen monimutkaisia bayesiläisiä tilastollisia malleja.