Johdatus tilastolliseen päättelyyn
kevät 2009
(korjattu versio kevään 2010 kurssia varten)

Elja Arjas Jukka Sirén

16. maaliskuuta 2010

Kurssin tavoitteista ja sisällöstä

Tällä luentokurssilla käsitellään *tilastollista päättelyä* eli *tilastollista inferenssiä* (engl. *statistical inference*). Aiheen käsittely eroaa — sekä sisällön että esitystavan puolesta — melkoisen paljon useimmista tilastotieteen peruskursseista. Kurssilla painotetaan tavanomaista enemmän tilastollisen päättelyn yhteydessä käytettyjä käsitteitä sekä niiden sisältöjä ja tulkintoja, kiinnittäen sitten vastaavasti vähemmän huomiota moniin käytännön soveltavassa tutkimuksessa tavallisiin tilastollisiin menetelmiin (esimerkiksi erilaisiin tilastollisiin merkitsevyystesteihin). Toisaalta luennoilla pyritään noudattamaan käytäntöä, jossa kaikkiin teoreettisiin käsitteisiin johdutaan sopivien havaintoesimerkkien kautta.

Kurssin seuraaminen ei edellytä aiempia tietoja tilastotieteestä ja sen on ajateltu sopivan hyvin kuunneltavaksi ja suoritettavaksi ensimmäisen opiskeluvuoden aikana. Kuitenkin kurssi perustuu lähes kauttaaltaan todennäköisyyslaskentaan, joten tässä suhteessa kohtuullinen matemaattinen sujuvuus — sekä varsinaisen kalkyylin että todennäköisyyslaskennassa esiintyvien keskeisten käsitteiden tulkintojen osalta — helpottaa kurssin seuraamista suuresti.

Kurssin luentojen sijoittaminen kevään jälkimmäiselle periodille suoraan kurssin *Johdatus todennäköisyyslaskentaan* jatkoksi voidaan tulkita ohjeeksi suorittaa nämä kaksi kurssia peräkkäin. Myös mm. tilastotieteen tutkintovaatimuksissa nämä kaksi jaksoa esiintyvät yhtenä kurssina *Johdatus todennäköisyyslaskentaan ja tilastotieteeseen*.

Tilastollinen päättely voidaan karkeasti ottaen jakaa

- parametriestimointiin, jossa pyritään arvioimaan tilastollisen mallin sisältämän tuntemattoman malliparametrin arvoa tehtyjen havaintojen avulla;
- ennusteongelman tarkasteluun, jossa aiempien havaintojen avulla pyritään ennustamaan havaintosarjan tulevia — ja siksi vielä tuntemattomia — arvoja;
- tilastollisten hypoteesien testaukseen, jossa tutkitaan mallia koskevien väitteiden paikkansapitävyyttä havaintoaineiston valossa, sekä
- mallinarviointiin (diagnostiikkaan), jossa koetetaan arvioida käytetyn mallin sopivuutta kuvaamaan tarkasteltua ilmiötä ja selittämään siitä tehtyjä havaintoja.

Tällä luentokurssilla käsitellään jonkin verran kaikkia näitä aiheita, mutta pääpaino on selvästi parametriestimoinnissa.

Sisältö

Luku 1

Todennäköisyyslaskennasta tilastotieteeseen

1.1 Tilastollinen malli

Johdatteleva esimerkki. Oletamme, että kulhossa on samankokoisia ja samasta materiaalista valmistettuja valkoisia ja mustia palloja yhteensä N kappaletta. Merkitään valkoisten pallojen lukumäärää $K = \#\{\text{valkoiset pallot}\}$, jolloin luonnollisesti mustia palloja on $N - K$ kappaletta. Ajatellaan sitten vielä, että jokaisen pallon kylkeen on maalattu jokin luvuista $1, 2, \ldots, N$, kaikkiin palloihin eri luku, jolloin ne voidaan erottaa toisistaan yksinkertaisesti katsomalla. Jos kulhosta nostetaan sokkona yksi pallo, on luonnollisen symmetrian perusteella järkevää ajatella, että kaikkien pallojen todennäköisyys tulla poimituksi on sama. Tällöin siis asetetaan todennäköisyyden arvoksi

$$P(\text{nostettu pallo on numeroitu numerolla } j) = \frac{1}{N}, \qquad j = 1, 2, \ldots, N.$$

Jotta todennäköisyydellä olisi määritelty numeerinen arvo, täytyy luvun N luonnollisesti olla tunnettu. Edelleen, jos myös valkoisten pallojen lukumäärä K on tunnettu, saadaan todennäköisyyden yhteenlaskuominaisuuden perusteella välittömästi tulos

$$P(\text{nostettu pallo on valkoinen}) = \frac{K}{N}.$$

Kurssilla *Johdatus todennäköisyyslaskentaan* on osoitettu, kuinka tämän yksinkertaisen todennäköisyysmallin avulla voidaan helposti käsitellä myös useampia nostoja käsittäviä havaintosarjoja. Määrittelemme nyt satunnaismuuttujan X_i seuraavalla tavalla:

$$\begin{aligned} X_i &= 1, \quad \text{jos } i\text{:nnellä nostolla saadaan valkoinen pallo, ja} \\ X_i &= 0, \quad \text{jos } i\text{:nnellä nostolla saaadaan musta pallo.} \end{aligned}$$

Silloin voimme kirjoittaa $P(X_1 = 1) = K/N$ ja $P(X_1 = 0) = 1 - K/N$, eli nämä tulokset yhdistämällä lyhyesti

$$P(X_1 = x_1) = \left(\frac{K}{N}\right)^{x_1} \left(1 - \frac{K}{N}\right)^{1-x_1}, \qquad x_1 = 0, 1.$$

Jos pallo kunkin noston jälkeen palautetaan kulhoon ja kulhon sisältöä tämän jälkeen sekoitetaan perusteellisesti, on järkevää olettaa peräkkäisten nostojen tulokset *riippumattomiksi*. Ne voidaan myös ajatella saman kokeen toistoiksi, joissa tietyn tuloksen todennäköisyys säilyy samana riippumatta siitä, monennestako nostokerrasta on kysymys.

Tällöin siis i:nnen noston tuloksen todennäköisyys on sama riippumatta siitä, mitä tätä aikaisemmilla nostoilla saatiin, ja se on myös sama kuin saman tuloksen todennäköisyys heti ensimmäisessä nostossa. Kaavan muodossa tämä voidaan kirjoittaa yhtäsuuruutena

$$P(X_i = x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) = P(X_i = x_i) = \left(\frac{K}{N}\right)^{x_i} \left(1 - \frac{K}{N}\right)^{1-x_i}.$$

Jos nyt tarkastellaan samanaikaisesti n:ää nostoa ja niissä saatuja tuloksia, saamme tällä perusteella todennäköisyyslaskennan kursseilta tutun tuloksen

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, X_n = x_n) \\ &\quad = P(X_1 = x_1)\, P(X_2 = x_2 | X_1 = x_1) \cdots \\ &\qquad\qquad P(X_n = x_n | X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \qquad (1.1) \\ &\quad = \left(\frac{K}{N}\right)^{T(\boldsymbol{x}_n)} \left(1 - \frac{K}{N}\right)^{n - T(\boldsymbol{x}_n)}, \end{aligned}$$

missä $T(\boldsymbol{x}_n) = \sum_{i=1}^{n} x_i$ on n:llä nostolla yhteensä saatu valkoisten pallojen lukumäärä. (Tässä on käytetty vektorimerkintää $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$.) Huomaa, että tämä todennäköisyys riippuu havaintotuloksista $x_1, x_2, \ldots, x_n$ vain niiden summan $T(\boldsymbol{x}_n)$ kautta.

Harjoitustehtävä 1.1 *Tarkastele edellä kuvattua nostokoetta, mutta nyt olettaen, ettei kerran nostettua palloa palauteta kulhoon. Johda tässä tilanteessa (edellä esitettyä mallia seuraten, mutta nyt luopuen ilmeisestikin paikkansapitämättömästä riippumattomuusoletuksesta) todennäköisyys $P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ tapauksessa $x_1 = 1$, $x_2 = 0$, $x_3 = 1$. Näytä, että tämä todennäköisyys on sama riippumatta siitä, missä järjestyksessä koetulos "kaksi valkoista palloa ja yksi musta pallo" saadaan. Päteekö sama myös muille mahdollisille koetuloksille? Palauta myös mieleesi hypergeometrisen jakauman käsite todennäköisyyslaskennan kurssilta.*

Lukuja K ja N voidaan kutsua satunnaismuuttujan (havaintovektorin)

$$\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$$

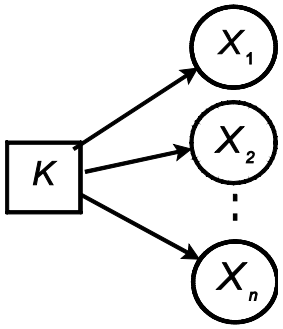
(todennäköisyys)jakauman *parametreiksi*. Oletamme nyt jatkossa yksinkertaisuuden vuoksi, että N on kiinteänä pysyvä tunnettu luku, jolloin parametriksi jää kulhossa olevien valkoisten pallojen lukumäärä K. Joskus tämä todennäköisyyksien riippuvuus parametrin arvosta kirjoitetaan selvyyden vuoksi näkyviin, esimerkiksi varustamalla todennäköisyydet niitä vastaavilla alaindekseillä, tässä siis P_K. Jakaumaperhettä

$$\boldsymbol{M} = \{P_K; 1 \leq K \leq N\}$$

kutsutaan *tilastolliseksi malliksi*. Tätä mallia voidaan havainnollistaa kuvalla 1.1.

Näin määritellyn mallin avulla, kun sen parametrin K arvo on kiinnitetty, voidaan nyt määrittää (yksinkertaisina summina kaavan (1.1) mukaisista todennäköisyyksistä) todennäköisyydet mielivaltaisille äärellisen pituisiin havaintosarjoihin liitettäville tapahtumille.

Todennäköisyyslaskennan ns. laajennuslauseen perusteella todennäköisyysjakaumaa P_K voidaan edelleen laajentaa siten, että sillä voidaan kuvata esimerkiksi valkoisten pallojen suhteellisen osuuden n:ssä nostossa (ts. muuttujien $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$) käyttäytymistä silloinkin, kun havaintojen lukumäärän n annetaan kasvaa rajatta. Ns. *suurten lukujen lain* (joka on todistettu tähänkin esimerkkitapaukseen sopivin oletuksin luentokurssilla *Johdatus todennäköisyyslaskentaan*) perusteella tiedetään, että ne suppenevat (jakauman P_K suhteen) kohti muuttujien X_i odotusarvoa $E_K(X_i) = K/N$.



Kuva 1.1: Tilastollinen malli $\boldsymbol{M}$.

Pohdintaa. Yleisesti ottaen voidaan sanoa, että tilastollisia malleja käytetään kuvaamaan havaittuja ja kohteeltaan rajattuja reaalimaailman ilmiötä, tulkiten malli tällöin eräänlaiseksi *dataa generoivaksi mekanismiksi*. Ajatuksena on tällöin, että tarkastellusta ilmiöstä konkreettisesti mittauksin tai muulla tavalla saadut havainnot voitaisiin ymmärtää ikään kuin ne olisi saatu mallin perusteella suoritetun tietokonesimulaation avulla. Kun näin toimitaan, pyritään tietenkin vastaavuuteen, jossa mallin antama kuvaus olisi tehtävän kannalta riittävän osuva ja jossa se koetaan jollakin tavalla hyödylliseksi. Tällöin tutkittavan ilmiön ominaispiirteitä ja käsitteitä voidaan ikään kuin ripustaa ilmiötä kuvaavan mallin varaan. Tämä tarjoaa luonnollisen perustan ilmiöön liittyvälle käsitteenmuodostukselle (vrt. esimerkiksi massan, voiman ja nopeuden käsitteet mekaniikassa), antaen samalla mahdollisuuden käyttää hyväksi matemaattisen analyysin tarjoamia keinoja "laskea kaavoilla" sekä myös mahdollisuutta suorittaa numeerisia laskutoimituksia. On kuitenkin huomattava, että matemaattinen tai tilastollinen malli ei yksinkertaisimmassakaan tapauksessa ole sama kuin tarkasteltava fysikaalisen maailman ilmiö: edellisessä esimerkissäkin varsinainen dataa generoiva mekanismi on kulho, siinä olevat pallot sekä henkilö, joka poimii palloja kulhosta.

Tilastollisista malleista ei useimmiten ole välitöntä konkreettista hyötyä, ellei niiden sisältämien parametrien arvoista ole käytettävissä edes jonkinlaista numeerista arviota. Huomaa kuitenkin, että käsite *parametrin oikea arvo* — jolla edellä olevassa esimerkissä on selkeästi määritelty konkreettinen vastine — voi joskus olla ongelmallisempi määriteltävä. Koska parametri on malliin liittyvä käsite, kirjaimellisesti tulkittuna sen oikea arvo voi viitata vain johonkin tiettyyn (esim. tietokonesimulaatiossa käytettyyn) malliin ja tässä yhteydessä valittuun parametrin arvoon. Tässä suhteessa on avuksi, jos parametrille voidaan esittää jokin operationaalinen määritelmä, esimerkiksi ajattelemalla jotakin ainakin periaatteessa mahdollista ideaaliolosuhteissa suoritettavaa mittausta.

1.2 Parametriestimointi uskottavuusfunktion perusteella

Johdatteleva esimerkki (jatkoa). Pohdimme nyt tilastolliselle päättelylle tyypillistä *käänteistä ongelmaa*: Oletetaan, että olemme tehneet havainnon $\boldsymbol{X}_n = \boldsymbol{x}_n$ toistamalla yllä esitettyä nostokoetta n kertaa. Oletamme kuitenkin, ettemme tunne valkoisten pallojen lukumäärää K. Mitä tehtyjen havaintojen avulla voidaan päätellä K:n arvosta? Tällaista ongelmaa kutsutaan *tilastolliseksi parametriestimoinniksi* — tai lyhemmin tilastolliseksi

estimoinniksi.

On selvää, että jo yhden pallon nostaminen kulhosta ja sen värin toteaminen antaa jotakin tietoa tuntemattomasta parametrista K. Erityisesti, jos pallo on valkoinen, ts. $X_1 = 1$, täytyy olla $K > 0$. Vastaavasti tapauksessa $X_1 = 0$ voidaan heti päätellä, että $K < N$. Lähellä on myös seuraava ajatus: Jos heti ensimmäinen käteen sattuva pallo on valkoinen (musta), ehkä niitä on kulhossa enemmän kuin mustia (valkoisia). Toisaalta, vaikka kaikki nostetut pallot olisivat valkoisia (mustia), emme voi olla täysin varmoja siitä, ettei niitä ole kulhossa kuin yksi ainoa, joka sitten osui käteen joka kerralla. Kuitenkin näemme kaavasta (1.1), että tällaisen pelkkiä valkoisia palloja sisältävän havaintosarjan todennäköisyys on $P_K(X_1 = 1, X_2 = 1, \ldots, X_n = 1) = (K/N)^n$, joka lähestyy nollaa otoskoon n kasvaessa kun $K < N$. Jossakin — toistaiseksi vielä tarkemmin määrittelemättömässä mielessä — jotkut K:n arvot vaikuttavat ilmeisesti tietyn otoksen valossa uskottavammilta kuin toiset. Toisaalta esitettyihin arvioihin sisältyvä epävarmuus, joka ei koskaan kokonaan poistu (ellei kulhoon saa katsoa), vähenee kuitenkin otoskoon kasvaessa.

Yleisesti voimme todeta, että malliparametrin arvon muuttuessa myös tiettyyn havaintoon $\boldsymbol{X}_n = \boldsymbol{x}_n$ liitettävä todennäköisyys muuttuu. Tätä todennäköisyyden riippuvuutta parametrin arvosta voidaan korostaa merkinnällä

$$L(K) = P_K(\boldsymbol{X}_n = \boldsymbol{x}_n), \tag{1.2}$$

jossa ao. todennäköisyys ymmärretään nimenomaisesti parametrin K funktioksi. Tilastotieteessä näin määriteltyä funktiota kutsutaan (havaintotulosta $\boldsymbol{x}_n$ vastaavaksi) *uskottavuusfunktioksi* (engl. *likelihood function*).

Nyt on varsin lähellä ajatus liittää estimointiongelman ratkaisu suoraan uskottavuusfunktion saamiin arvoihin. Tämä voidaan lausua periaatteena: Parametrin K arvo on (havainnon $\boldsymbol{x}_n$ valossa) sitä uskottavampi kuta suurempi uskottavuusfunktion $L(K)$ arvo on. Tässä mielessä "uskottavin arvio" tuntemattomalle parametrin arvolle K saadaan etsimällä se arvo, joka maksimoi lausekkeen (1.2). Näin määriteltyä lukua kutsutaan (havaintotulosta $\boldsymbol{x}_n$ vastaavaksi) *suurimman uskottavuuden estimaatiksi* (engl. *maximum likelihood estimate*), ja sitä merkitään tavallisesti symbolilla $\hat{K}$.

Näin määritelty parametriestimaatti on havaintotuloksen $\boldsymbol{x}_n$ funktio. Jos halutaan korostaa estimaatin funktioluonnetta, ts. sen riippuvuutta tästä tuloksesta, korvataan sana *estimaatti* usein sanalla *estimaattori* (engl. *estimator*). Sillä on luonnollisesti kiinteä arvo sen jälkeen kun havainto $\boldsymbol{X}_n = \boldsymbol{x}_n$ on tehty. Toisaalta, niin kauan kuin pidämme havaintoa satunnaismuuttujana $\boldsymbol{X}_n$, on myös uskottavuusfunktion (1.2) maksimoiva parametrin K arvo satunnaismuuttuja. Selvyyden vuoksi sitä voitaisiin merkitä $\hat{K} = \hat{K}(\boldsymbol{X}_n)$.

Johdamme nyt em. esimerkkitapauksessa havaintotulosta $\boldsymbol{X}_n = \boldsymbol{x}_n$ vastaavan suurimman uskottavuuden estimaattorin $\hat{K}(\boldsymbol{x}_n)$ lausekkeen. Kaavojen (1.1) ja (1.2) perusteella voimme aluksi kirjoittaa uskottavuusfunktion lausekkeeksi

$$\begin{aligned} L(K) &= P_K(\boldsymbol{X}_n = \boldsymbol{x}_n) \\ &= P_K(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \left(\frac{K}{N}\right)^{T(\boldsymbol{x}_n)} \left(1 - \frac{K}{N}\right)^{n - T(\boldsymbol{x}_n)}, \end{aligned}$$

missä $T(\boldsymbol{x}_n) = \sum_{i=1}^{n} x_i$. Suurimman uskottavuuden estimaattorin määritelmäksi tulee siis

$$\hat{K} = \hat{K}(\boldsymbol{x}_n) = \arg\max_K \left(\frac{K}{N}\right)^{T(\boldsymbol{x}_n)} \left(1 - \frac{K}{N}\right)^{n - T(\boldsymbol{x}_n)}.$$

Tehtävän ratkaisua helpottaa, jos kirjoitamme parametrin K muotoon $K = \theta N$ ja ajattelemme aluksi θ:n (ja siten myös K:n) jatkuvaksi muuttujaksi, jolloin uskottavuusfunktion lauseketta voidaan maksimikohdan määrittämiseksi derivoida. Samalla uskottavuusfunktio tulkitaan θ:n funktioksi. Tehtävä helpottuu edelleen hieman, jos derivointi kohdistetaan $L(\theta)$:n asemesta sen logaritmiin $\log L(\theta)$. Derivaatan nollakohdaksi (ja siis parametrin θ suurimman uskottavuuden estimaatiksi) saadaan tällöin koetulosten aritmeettinen keskiarvo $\frac{1}{n}\sum_{i=1}^{n} x_i$ eli valkoisten pallojen suhteellinen osuus n:n noston koesarjassa. Määritelmän $K = \theta N$ perusteella voidaan sitten todeta, että tämä tulos kerrottuna luvulla N voidaan tulkita suoraan myös alkuperäisen parametrin K suurimman uskottavuuden estimaatiksi. Lopuksi, mikäli saatu tulos ei ole kokonaisluku ja sen kuitenkin haluttaisiin olevan sellainen, voidaan palata jälleen kokonaislukuarvoiseen parametriestimaattiin tarkastelemalla näin saatua estimaatin lukuarvoa $\frac{N}{n}\sum_{i=1}^{n} x_i$ välittömästi edeltävää ja sitä seuraavaa kokonaislukua ja valitsemalla niistä se, joka antaa uskottavuusfunktiolle suuremman arvon.

Harjoitustehtävä 1.2 *Suorita edellä käsitellyssä esimerkkitapauksessa tarvittavat laskut uskottavuusfunktion maksimikohdan määrittämiseksi.*

Harjoitustehtävä 1.3 *Osoita, että edellisessä esimerkkitapauksessa johdettu estimaattori* $\hat{K}(\boldsymbol{X}_n) = \frac{N}{n}\sum_{i=1}^{n} X_i$ *on odotusarvon mielessä* harhaton *(engl.* unbiased*) siinä mielessä, että* $E_K(\hat{K}(\boldsymbol{X}_n)) = K$ *kaikilla* K*:n arvoilla.*

Harjoitustehtävä 1.4 *Määritä myös tämän estimaattorin varianssi* $\mathrm{Var}_K(\hat{K}(\boldsymbol{X}_n))$. Opastus: *Todennäköisyyslaskennan kurssilla on johdettu tulos, jonka mukaan riippumattomille (yleisemmin korreloimattomille) satunnaismuuttujille pätee* $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

Edellä harjoitustehtävässä 1.4 johdetun estimaattorin varianssin lausekkeesta nähdään suoraan, että $\mathrm{Var}_K(\hat{K}(\boldsymbol{X}_n)) \to 0$ kun $n \to \infty$. Kun muistetaan, että varianssi kuvaa eräällä tavalla todennäköisyysjakauman kuvaaman satunnaisuuden määrää, voidaan siis sanoa, että estimaattori $\hat{K}(\boldsymbol{X}_n)$ on otoskoon n kasvaessa *tarkentuva* (engl. *consistent*). Tämä ominaisuus voidaan kirjoittaa täsmällisesti (edellä mainitun *suurten lukujen lain* mallia noudattaen) seuraavalla tavalla: jokaisella $h > 0$ pätee, että

$$P_K(|\hat{K}(\boldsymbol{X}_n) - K| > h) \to 0, \qquad \text{kun } n \to \infty.$$

Pohdintaa. On huomattava, että tällainen estimaattorin $\hat{K}$ luonnehdinta tarkentuvaksi, samoin kuin harjoitustehtävässä 1.3 tarkasteltu harhattomuuskin, ovat nimenomaisesti satunnaismuuttujaksi tulkitun estimaattorin $\hat{K} = \hat{K}(\boldsymbol{X}_n)$ ominaisuuksia. Jos sen argumenttina oleva muuttuja $\boldsymbol{X}_n$ kiinnitetään johonkin tämän muuttujan havaittuun arvoon $\boldsymbol{x}_n$, myös estimaattori $\hat{K}(\boldsymbol{X}_n)$ saa luonnollisesti tätä vastaavan vakioarvon $\hat{K}(\boldsymbol{x}_n)$. Tarkentuvuusominaisuus — tai muut edellä tarkastellut estimaattorin jakaumaominaisuudet — eivät enää anna välitöntä tietoa siitä, kuinka hyvä tai tarkka arvio jonkin tietyn havaintotuloksen $\boldsymbol{x}_n$ perusteella määritetty luku $\hat{K}(\boldsymbol{x}_n)$ on tuntemattomalle parametrinarvolle K. Ei siis voida sanoa esimerkiksi, että suurimman uskottavuuden estimaattorin $\hat{K}(\boldsymbol{X}_n)$ saama havaittu arvo $\hat{K}(\boldsymbol{x}_n)$ olisi jossakin mielessä *todennäköisin* parametrin K arvo. Tämä johtuu yksinkertaisesti siitä, että tarkastellussa mallissa $\boldsymbol{M} = \{P_K; 1 \le K \le N\}$ todennäköisyydet P_K on liitetty vain muuttujien $\boldsymbol{X}_n$ saamiin arvoihin, eivät parametrin K arvoihin. Tämän mallin puitteissa ei ole olemassa todennäköisyyttä, jonka avulla voitaisiin luonnehtia parametrin K saamiin arvoihin liittyviä väitteitä.

1.3 Todennäköisyys uskomuksen asteena: parametri satunnaismuuttujana

Edellä esitetty pohdinta antaa aiheen kysyä, eikö estimointiongelmassa ilmenevää, parametrin "oikeaan arvoon" liittyvää, epävarmuutta kuitenkin olisi paikallaan kuvata käyttäen tähän todennäköisyyskäsitettä? Jos näin voitaisiin tehdä, olisi mahdollista ilmaista suoraan esimerkiksi mitkä parametrin arvot näyttävät havaintotulosten valossa todennäköisiltä ja mitkä taas vähemmän todennäköisiltä. Näin ei kuitenkaan voi menetellä suoraan, ellei mallimäärittelyä muuteta siten, että myös parametri siinä ymmärretään satunnaismuuttujaksi. Esimerkkitapauksessa siis valkoisten pallojen lukumäärä kulhossa K tulkittaisiin satunnaismuuttujaksi. Ennen kuin näin voidaan menetellä, on hyvä pohtia hieman sitä, mitä tällainen pallojen lukumäärän satunnaisuus voisi merkitä.

Voimme tietenkin ajatella, että tietyssä koetilanteessa kulhossa olevien pallojen todelliseen lukumäärään on johduttu jollakin fysikaalisella satunnaismekanismilla, arvonnalla tms. Tällöin siis ajatellaan, että lukumäärä K on ensiksi poimittu jostakin jakaumasta, jonka jälkeen sitten palloja ryhdytään poimimaan sellaisesta kulhosta, jonka sisältämistä N:stä pallosta K on valkoisia. Sen jälkeen kun lukumäärä K on kiinnitetty, yksittäisiä havaintoja (nostoja) X_i voidaan pitää riippumattomina. Jos merkitsemme edelleen valkoisten pallojen lukumäärää K:lla, mutta tulkitsemme sen satunnaismuuttujaksi sekä merkitsemme tämän muuttujan saamia arvoja k:lla, voimme todennäköisyyslaskennan kertolaskusääntöä käyttämällä kirjoittaa

$$P(K = k, \boldsymbol{X}_n = \boldsymbol{x}_n) = P(K = k)\, P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k). \tag{1.3}$$

Tämän tulon jälkimmäinen termi saa täsmälleen saman arvon kuin vastaava todennäköisyys kaavassa (1.1), ts.

$$P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k) = \left(\frac{k}{N}\right)^{T(\boldsymbol{x}_n)} \left(1 - \frac{k}{N}\right)^{n - T(\boldsymbol{x}_n)}.$$

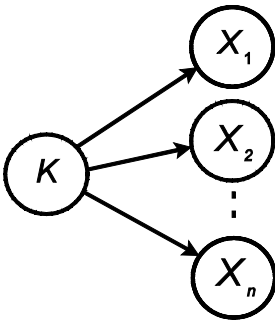
Ero on siinä, että nyt K tulkitaan satunnaismuuttujaksi, jonka saamiin arvoihin siis voidaan liittää todennäköisyyksiä, kun taas mallin $\boldsymbol{M} = \{P_K; 1 \leq K \leq N\}$ puitteissa tämä ei ollut sallittua. Toinen ero on siinä, että aiemmin tarkasteltu muuttujien X_i riippumattomuus oli todennäköisyyden P_K suhteen määritelty ominaisuus, kun taas todennäköisyyden $P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k)$ lausuminen tulomuodossa

$$P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k) = \prod_{i=1}^{n} P(X_i = x_i \mid K = k) = \left(\frac{k}{N}\right)^{T(\boldsymbol{x}_n)} \left(1 - \frac{k}{N}\right)^{n - T(\boldsymbol{x}_n)} \tag{1.4}$$

tulee ymmärtää tapahtumien $\{X_i = x_i\}$ *ehdollisena riippumattomuutena* todennäköisyyden P suhteen, kun ehtona on tapahtuma $\{K = k\}$.

Yhtälön (1.3) vasemmalla puolella esiintyvät todennäköisyydet määrittelevät muuttujien K ja $\boldsymbol{X}_n$ ns. *yhteisjakauman*. Toisaalta siihen liitettävät todennäköisyydet saadaan määritetyksi tämän yhtälön oikealta puolelta muuttuja kerrallaan, ts. tarkastelemalla ensin muuttujan K arvoihin ja sen jälkeen muuttujan $\boldsymbol{X}_n$ arvoihin liitettäviä todennäköisyyksiä. Myös tämän jakauman kohdalla voidaan käyttää termiä *tilastollinen malli*, jota nyt merkitsemme kirjaimella $\boldsymbol{M}^*$.

Tätä mallia voidaan havainnollistaa kuvalla 1.2. Vaikka tämä kuva onkin muuten samanlainen kuin kuva 1.1, tärkeä ero on siinä, että nyt K on satunnaismuuttuja ja siksi se



Kuva 1.2: Tilastollinen malli $\boldsymbol{M}^*$.

on piirretty ympyrän sisään. Tämä vastaa mallien $\boldsymbol{M}^*$ ja $\boldsymbol{M}$ määrittelyssä ollutta eroa: edellinen määrittelee muuttujien K ja $\boldsymbol{X}_n$ yhteisjakauman, jälkimmäinen taas muuttujan $\boldsymbol{X}_n$ jakaumien P_K muodostaman jakaumaperheen, jonka parametrina on K. On myös tärkeää huomata, ettei mallissa $\boldsymbol{M}^*$ päde muuttujien X_i riippumattomuus ilman muuttujan K arvoon liittyvää muotoa $\{K = k\}$ olevaa ehtoa. Yhtälöstä (1.4) ei siis voi jättää tähän liittyvää ehdollistamista pois. Intuitiivisesti tämä merkitsee sitä, että — ellei kulhossa olevien valkoisten pallojen lukumäärä K ole tiedossa — siitä saadaan kuitenkin epäsuoraa tietoa aiempien havaintojen $X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}$ perusteella. Näin ollen koesarjassa i:nteen nostoon liitettävä todennäköisyys

$$P(X_i = x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

ei yleensä ole sama kuin *reunatodennäköisyys* $P(X_i = x_i)$. Toisaalta kaikki reunatodennäköisyydet ovat keskenään yhtä suuria ja siten myös yhtä suuria kuin heti ensimmäiseen nostoon liitettävä todennäköisyys $P(X_1 = x_1)$, joka voidaan mallissa $\boldsymbol{M}^*$ laskea ns. *kokonaistodennäköisyyden* kaavaa käyttäen:

$$P(X_1 = x_1) = \sum_k P(K = k, X_1 = x_1) = \sum_k P(K = k)\, P(X_1 = x_1 \mid K = k). \qquad (1.5)$$

Pohdintaa. Voidaan kysyä, onko jonkinlaisen fysikaalisen arvontamekanismin olemassaolo välttämätön edellytys sille, että voisimme tässä yhteydessä puhua pallojen lukumäärästä satunnaismuuttujana ja sitten liittää sen arvoihin todennäköisyyksiä? Edellä kuvatussa tilanteessa, jossa palloja ryhdytään nostamaan kulhosta, siellä on tietenkin jokin määrä valkoisia palloja ja se on korkeintaan sama kuin pallojen kokonaismäärä kulhossa. Valkoisten ja mustien pallojen lukumäärät eivät peräkkäisillä nostoilla muutu, koska nostettu pallo palautetaan aina kulhoon takaisin ennen kuin tehdään uusi nosto. Niiden lukumääriä koskevan arviointiongelman tarkastelun kannalta ei näyttäisikään olevan olennaista se, onko valkoisten pallojen lukumäärä saatu esim. arpomalla, vaan se, tiedämmekö me — tai tiedänkö minä — montako valkoista palloa siellä on?

Suoraviivainen ratkaisu satunnaisuutta koskevaan kysymykseen voisi siis olla seuraava: *Pidän tarkasteltavaa suuretta satunnaismuuttujana, jos en tiedä mikä arvo sillä on.* Jos näin ajatellaan, todennäköisyydestä tulee ilmaus *uskomuksen asteesta* (engl. *degree of belief*).

Ns. Bayes-päättelyssä nojaudutaankin lähinnä *todennäköisyyden subjektiiviseen tulkintaan*, jonka mukaan on mahdollista (ja sallittua!) liittää todennäköisyyksiä myös mm. jonkin tarkastellun jakaumaperheen parametrin arvoihin. Jos näin menetellään, parametrin arvoihin liitettävät todennäköisyydet voidaan ymmärtää kvantitatiivisina ilmauksina subjektin (esim. "minun") epävarmuudesta koskien parametrin oikeaa arvoa. Eri henkilöt (ja myös sama henkilö eri aikoina, jos käytettävissä oleva informaatio muuttuu) voivat siten esittää erilaisia numeerisia arvioita saman tapahtuman todennäköisyyksistä. Tämän subjektiivisen tai "henkilökohtaisen" tulkinnan mukaan satunnaisuudessa ei välttämättä ole kysymys siitä, että tarkasteltavat tapahtumat (esimerkiksi satunnaismuuttujien saamat arvot) vaihtelisivat ajan mukana tai että ne määräytyisivät konkreettisesti vasta tarkasteluhetkeä seuraavassa tulevaisuudessa.

Ääritapauksessa voimme tarkastella esimerkiksi tiettyä historiallista tapahtumaa koskevan väitettä tai tieteellistä hypoteesia — joka sinänsä on joko tosi tai epätosi — samalla tavalla kuin "satunnaistapahtumaa", liittämällä väitteeseen H ja sille vastakkaiseen väitteeseen H^C todennäköisyydet $P(H)$ ja $P(H^C)$. Koska joko H tai H^C on joka tapauksessa oikea, oletamme luonnollisen ehdon $P(H)+P(H^C) = 1$. Jos olen melko varma siitä, että H on oikea, voin kuvata tätä antamalle $P(H)$:lle lukuarvon, joka on lähellä ykköstä. $P(H^C)$ on silloin lähellä nollaa. Jos taas pidän väitettä H^C melko varmasti oikeana, H on melko varmasti väärä, ja tätä voidaan jälleen kuvata vastaavien todennäköisyyksien $P(H)$ ja $P(H^C)$ lukuarvoilla. Jos olen "täysin epävarma" siitä, kumpi väite on oikea, voin kuvata tätä asettamalla $P(H) = P(H^C) = \frac{1}{2}$.

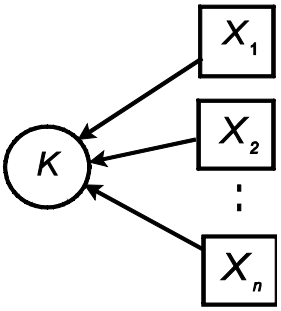
On tärkeää todeta, että vaikka todennäköisyyksien numeeriset arvot voivatkin vaihdella edellä kuvatulla tavalla, todennäköisyyslaskenta matemaattisena struktuurina pysyy samana riippumatta siitä, minkä tulkinnan valossa todennäköisyyksiä tarkastellaan. Sen sijaan, kun samaa kysymystä tarkastellaan tilastollisen päättelyn kannalta, joudutaan toteamaan, että todennäköisyyskäsitteen erilaiset tulkinnat voivat johtaa perusteellisesti erilaisiin menettelytapoihin ja ratkaisuihin. Näin tilastollisessa päättelyssä joudutaankin tekemään selvä ero kahden periaatteellisesti toisistaan eroavan paradigman, ns. *frekventistisen* ja *Bayes-paradigman* välillä.

1.4 Parametriestimointi käänteisenä ongelmana: Bayesin kaava

Palaamme nyt *käänteisen ongelman* tarkasteluun, ts. kysymykseen siitä, miten "oikeaa" (mutta havaitsijalle tuntematonta) kulhossa olevien pallojen lukumäärää K voitaisiin arvioida tehtyjen havaintojen perusteella. Mallin $\boldsymbol{M}^*$ puitteissa johdumme luonnollisella tavalla tehtävään, jossa pitäisi määrittää ehdolliset todennäköisyydet

$$P(K = k \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n), \qquad 0 \leq k \leq N. \tag{1.6}$$

Asiaa voidaan jälleen havainnollistaa seuraavalla kuvalla 1.3. Edelliseen kuvaan 1.2 verrattuna tässä on tapahtunut kaksi tärkeää muutosta: Muuttujat X_i on nyt piirretty laatikoiden sisään, vastaten kaavan (1.6) mukaista tilannetta, jossa niiden arvot ovat havaintojen perusteella tunnettuja. Lisäksi nuolten suunta on muuttunut, sillä nyt tarkastelemme vaihtoehtoihin $\{K = k\}$ liitettävien todennäköisyyksien riippuvuutta muuttujista X_i, eikä päinvastoin kuten edellä. Tätä vastaavasti voidaan sanoa, että kaavoissa (1.4) ja (1.6) muuttujat K ja $\boldsymbol{X}_n$ ovat vaihtaneet paikkoja. Näiden kahden ehdollisen todennäköisyyden, joissa tarkasteltava tapahtuma ja ehtotapahtuma vaihtavat



Kuva 1.3:

järjestystä, välinen yhteys voidaan ilmaista tunnetun (ja mm. luentokurssilla *Johdatus todennäköisyyslaskentaan* käsitellyn) *Bayesin kaavan* avulla, jota nyt sovelletaan.

Bayesin kaava esitetään todennäköisyyslaskennassa tavallisesti muodossa

$$P(B_k|A) = \frac{P(B_k \text{ ja } A)}{P(A)} = \frac{P(B_k)\,P(A|B_k)}{\sum_i P(B_i)\,P(A|B_i)}, \tag{1.7}$$

missä A on mielivaltainen tapahtuma ja tapahtumat B_i muodostavat jonkin todennäköisyyskentän perusjoukon Ω (*varman tapahtuman*) osituksen, ts. $\Omega = \bigcup_i B_i$ ja osat B_i ovat pistevieraita. Kaavassa (1.7) ensimmäinen yhtäsuuruus vastaa suoraan ehdollisen todennäköisyyden määritelmää, toinen taas saadaan käyttämällä todennäköisyyslaskennan kertolaskukaavaa sekä lisäksi nimittäjässä kokonaistodennäköisyyden kaavaa. Kun merkitään edellä $A = \{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \{\boldsymbol{X}_n = \boldsymbol{x}_n\}$ ja $B_k = \{K = k\}$, saadaan siis tulos

$$P(K = k|\boldsymbol{X}_n = \boldsymbol{x}_n) = \frac{P(K = k, \boldsymbol{X}_n = \boldsymbol{x}_n)}{P(\boldsymbol{X}_n = \boldsymbol{x}_n)} = \frac{P(K = k)P(\boldsymbol{X}_n = \boldsymbol{x}_n|K = k)}{\sum_i P(K = i)P(\boldsymbol{X}_n = \boldsymbol{x}_n|K = i)}. \tag{1.8}$$

Oikealla puolella nimittäjänä oleva summa voidaan ymmärtää normeerausvakioksi, joka ei riipu argumentista k ja joka vain takaa sen, että tulos on todennäköisyysjakauma satunnaismuuttujan K arvojoukossa $\{0, 1, \ldots, N\}$. Toisin sanoen, kun lasketaan oikealla puolella summa yli kaikkien k:n arvojen, pitää tuloksen olla = 1. Tästä syystä kaava (1.8) kirjoitetaan usein lyhyemmässä muodossa

$$P(K = k|\boldsymbol{X}_n = \boldsymbol{x}_n) \propto P(K = k)P(\boldsymbol{X}_n = \boldsymbol{x}_n|K = k), \tag{1.9}$$

missä merkintä "$\propto$" tarkoittaa verrannollisuutta k:n suhteen (ts. yhtäsuuruutta mahdollista k:sta riippumatonta vakiotekijää lukuun ottamatta). Kaavan (1.9) oikealta puolelta puuttuva normeeraustekijä voidaan luonnollisesti aina määrittää kaavasta jälkikäteen, mikäli niin halutaan, yksinkertaisesti summaamalla.

Kaavan (1.9) oikean puolen ensimmäistä tekijää $P(K = k)$ kutsutaan usein tapahtuman $\{K = k\}$ *prioritodennäköisyydeksi*. Tällä termillä pyritään ilmaisemaan, että kysymys on tällöin eräänlaisesta ennakkoarviosta, jonka pohjana ei välttämättä ole aiempien toistokokeiden tuottamia havaintoja muuttujista X_i. Prioritodennäköisyydet voidaan periaatteessa valita millä tavalla tahansa, mutta käytännön tilanteissa on tietenkin

paikallaan pyrkiä määrittämään ne siten, että ne vastaisivat läheisesti todellisia ennakkokäsityksiä. Oikean puolen toinen tekijä puolestaan on edellä kaavalla (1.4) määritellyn uskottavuusfunktion arvo pisteessä k. Kaavojen (1.8) ja (1.9) sisältämää tulosta voidaan nyt luonnehtia siten, että tapahtumaan $\{K = k\}$ ennakkoon liitettyä prioritodennäköisyyttä on päivitetty havaintojen $\boldsymbol{X}_n = \boldsymbol{x}_n$ perusteella ja näiden kaavojen osoittamalla tavalla. Tulosta $P(K = k|\boldsymbol{X}_n = \boldsymbol{x}_n)$ kutsutaan tapahtuman $\{K = k\}$ *posteriori-todennäköisyydeksi*. Tarkasteltaessa näitä todennäköisyyksiä satunnaismuuttujan K koko arvojoukossa $\{0, 1, \ldots, N\}$ puhutaan vastaavasti tämän muuttujan *priori-* ja *posteriori-jakaumasta*.

Esimerkki 1.1 *Eräällä laboratoriotestillä pyritään turvallisuussyistä selvittämään sitä, ovatko verta luovuttamaan tulleet henkilöt mahdollisesti HIV:n kantajia. Käytännössä tämä tapahtuu tutkimalla kaikkien verenluovuttajien osalta, sisältääkö veri HIV:n vasta-aineita. Oletamme testin herkkyyden olevan* 0.997*, ts. testitulos on positiivinen tällä todennäköisyydellä, jos luovuttajalla todella on veressään vasta-aineita. Toisaalta oletetaan, että testi antaa todennäköisyydellä* 0.015 *(väärän) positiivisen tuloksen silloinkin, kun vasta-aineita ei oikeasti ole. Oletamme vielä, että HIV-vasta-aineita on väestössä noin yhdellä tuhannesta ja että verenluovuttajat eivät käytännössä ole valikoitu otos väestöstä. (Tämä oletus voi käytännössä olla hyvin epärealistinen!) Millä todennäköisyydellä satunnaisesti valittu luovuttaja antaa positiivisen näytteen? Millä todennäköisyydellä positiivisen testituloksen saaneen henkilön veri sisältää oikeasti HIV-vasta-aineita?*

Ratkaisu. Merkitään satunnaisesti valittua henkilöä koskevia väitteitä seuraavasti

$$\begin{aligned} A &: \text{Positiivinen testitulos;} \\ B_1 &: \text{Veressä on HIV-vasta-aineita ja} \\ B_2 &: \text{Veressä ei ole HIV-vasta-aineita.} \end{aligned}$$

Positiivisen testituloksen eli tapahtuman A todennäköisyys saadaan laskettua kokonaistodennäköisyyden kaavalla:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \\ &= 0.997 \times 0.001 + 0.015 \times 0.999 \approx 0.016. \end{aligned}$$

Vasta-aineiden todennäköisyys veressä positiivisen testituloksen saaneella henkilöllä voidaan laskea Bayesin kaavan avulla:

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(A)} \approx \frac{0.997 \times 0.001}{0.016} \approx 0.0624.$$

Positiivisen testituloksen saaneella henkilöllä ei siis luultavimmin ole veressään HIV-vasta-aineita.

Esimerkki 1.2 *Naapuriin on muuttanut perhe, josta sinulle on kerrottu, että heillä on kaksi lasta. Tarkastele seuraavia tilanteita:*

(a) Haluat tutustua heihin hieman lähemmin ja soitat naapurin ovikelloa, jolloin avaamaan tulee noin kymmenvuotias poika, arvatenkin toinen perheen lapsista.

(b) Talonmies, joka on nähnyt perheen molemmat lapset, kertoo sinulle hieman arvoituksellisesti, että "ainakin toinen lapsista on poika".

Vastaa kummassakin tapauksessa kysymykseen: Mikä on todennäköisyys, että toinenkin perheen lapsista on poika? Oletamme tässä, että syntyvän lapsen sukupuoli määräytyy kullakin kerralla riippumattomasti ja että se on kummallekin sukupuolelle $\frac{1}{2}$.

Ratkaisu. Käytetään eri vaihtoehdoille seuraavia merkintöjä

B_1 : Molemmat lapset ovat poikia;
B_2 : Molemmat lapset ovat tyttöjä;
B_3 : Toinen lapsi on poika ja toinen tyttö.

Näiden *a priori* todennäköisyydet ovat $P(B_1) = P(B_2) = \frac{1}{4}$ ja $P(B_3) = \frac{1}{2}$.

(a) Merkitään A:lla tilanteen (a) mukaista havaintoa, että toinen lapsista on poika, ja tutkitaan A:n ehdollisia todennäköisyyksiä. Ehdolla B_1 A:n todennäköisyys on 1 ja ehdolla B_2 sen todennäköisyys on 0. Ehdolla B_3 A:n ehdollinen todennäköisyys on $\frac{1}{2}$, koska avaamaan tullut lapsi ajatellaan arvotuksi kahden lapsen joukosta. Bayesin kaavaa käyttäen tapahtuman B_1 ehdolliseksi todennäköisyydeksi ehdolla A saadaan

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{\sum_i P(B_i)P(A|B_i)} = \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{2}.$$

(b) Merkitään C:llä tilanteen (b) mukaista havaintoa, että ainakin toinen lapsista on poika. C:n ehdollinen todennäköisyys sekä ehdolla B_1 että ehdolla B_3 on 1, koska väite pitää paikkansa molemmissa tapauksissa. Ehdolla B_2 tämä todennäköisyys on 0. Bayesin kaavaa käyttäen saadaan tapahtuman B_1 ehdolliseksi todennäköisyydeksi ehdolla C

$$P(B_1|C) = \frac{P(B_1)P(C|B_1)}{\sum_i P(B_i)P(C|B_i)} = \frac{\frac{1}{4} \times 1}{\frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{2} \times 1} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}.$$

Harjoitustehtävä 1.5 *Mikä ehto todennäköisyyksien $P(A)$ ja $P(B)$ tulee toteuttaa, jotta $P(A|B) = P(B|A)$? (Tenttitehtävä 25.5.04)*

Harjoitustehtävä 1.6 *Kulhossa on 3 palloa, joista K valkoisia ja loput mustia. Et kuitenkaan tiedä, kuinka suuri K on, joten katsot voivasi kuvata tilannetta antamalla kaikille vaihtoehdoille $\{K = 0\}$, $\{K = 1\}$, $\{K = 2\}$ ja $\{K = 3\}$ saman prioritodennäköisyyden. Saat luvan nostaa kulhosta "sokkona" kaksi palloa siten, että ennen toista nostoa ensimmäisenä nostettu pallo palautetaan takaisin kulhoon, jonka jälkeen kulhoa ravistetaan perusteellisesti. Saat tuloksena kummallakin kerralla valkoisen pallon. Määritä tällä perusteella K:n posteriorijakauma. (Tenttitehtävä 16.11.05)*

Harjoitustehtävä 1.7 *(*)*[1] *Ajatellaan, että havainnot $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ tehdään kahdessa jaksossa, esim. kahden päivän aikana, jolloin ensimmäisenä päivänä havaitaan $X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i$ ja toisena $X_{i+1} = x_{i+1}, \ldots, X_n = x_n$. Ajatellaan sitten, että K:n posteriorijakauma määritetään näiden havaintojen perusteella kahdella vaihtoehtoisella tavalla:*

[1] Tähdellä (*) merkittyjen tehtävien ratkaisusta saa kaksinkertaisen määrän laskuharjoituspisteitä.

(i) Aluksi prioritodennäköisyydet $P(K = k)$ päivitetään ensimmäisen jakson havaintojen avulla posterioritodennäköisyyksiksi $P(K = k|X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i)$. Tämän jälkeen näin saatu jakauma tulkitaan priorijakaumaksi toisen jakson mittauksille ja päivitetään jälleen posteriorijakaumaksi, mutta nyt havaintojen $X_{i+1} = x_{i+1}, \ldots, X_n = x_n$ avulla.

(ii) Päivitys suoritetaan yhdellä kertaa, käyttäen kaikkia saatuja havaintoja.

Näytä, että molemmilla menetelmillä tulokseksi saadaan sama posteriorijakauma.

[Opastus: Merkintöjen yksinkertaistamiseksi voit aluksi määritellä tapahtumat $A = \{K = k\}$, $B = \{X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i\}$ ja $C = \{X_{i+1} = x_{i+1}, \ldots, X_n = x_n\}$. Nyt voit tilanteessa (i) soveltaa Bayesin kaavaa kahdesti, aluksi päivittäen prioritodennäköisyyden $P(A)$ B:n perusteella ehdolliseksi todennäköisyydeksi $P(A|B)$ ja sitten päivittäen tämän edelleen C:n perusteella ehdolliseksi todennäköisyydeksi $P(A|B$ ja $C)$. Tapauksessa (ii) tehdään vain yksi $P(A)$:n päivitys, nyt tapahtuman "B ja C" perusteella. On siis näytettävä, että molemmilla tavoilla saadaan sama ehdollinen todennäköisyys $P(A|B$ ja $C)$.]

Harjoitustehtävä 1.8 *(*) Televisio-show'n "Let's make a deal" isäntä Monty Hall esittelee kilpailijalle kolme koppia, jotka on numeroitu yhdestä kolmeen. Kopit ovat verhojen takana. Yhdessä niistä on tuliterä auto. Jos kilpailija arvaa, minkä verhon takana auto on, hän saa sen. Kilpailija valitsee sattumanvaraisesti kopin numero yksi. Isäntä, joka tietää jo, missä kopissa auto on, avaa nyt kopin numero kaksi. Hän näyttää, että se on tyhjä ja tekee tarjouksen. "Haluatko nyt pitää kiinni ensimmäisestä arvauksestasi, eli kopista numero yksi, vai vaihtaa sen koppiin numero kolme?"*

Kuvittele itsesi kilpailijan paikalle. Miten toimit? Millä perusteella?

[Tehtävä on poimittu kirjasta Hans Christian von Baeyer: Informaatio (suom. Timo Paukku), Terra Cognita 2005. Kirjassa on muutakin mielenkiintoista luettavaa tämän kurssin sisältöä ajatellen, mm. luku 9 "Todennäköisen päättelyä: Kuinka todennäköisyys mittaa informaatiota" ja luku 12 "Satunnaisuus: Informaation kääntöpuoli". Tehtävää käsittelee myös Wikipedian sivu http://en.wikipedia.org/wiki/Monty_Hall_problem.]

1.5 Estimoinnista ennustamiseen: mikä pallo seuraavaksi?

Jatkamme edellä käsitellyn esimerkin tarkastelua, mutta vaihtamalla siinä hieman näkökulmaa ja siirtymällä valkoisten pallojen tuntemattoman lukumäärän K arvioinnista tehtävään, jossa pyrimme ennustamaan myöhemmin suoritettavan noston tulosta. Tarkastelemme siis tilannetta, jossa muuttujan K arvoa ei tunneta, mutta jossa meillä on käytettävissä aiemmista n:stä nostosta saadut havainnot $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Millä todennäköisyydellä seuraava nosto tuottaa valkoisen (mustan) pallon?

Tarkastelemme tätä kysymystä ensin mallin $\boldsymbol{M}$ puitteissa. Jos meillä olisi käytettävissä tieto valkoisten pallojen lukumäärästä K, ts. tiedämme että $K = k$, olisi ilmeinen vastaus esitettyyn kysymykseen: "Todennäköisyys sille, että seuraavalla nostolla saadaan valkoinen pallo, on k/N". Koska kuitenkin täsmällinen tieto K:n arvosta puuttuu, yksi luonteva mahdollisuus on korvata se osamäärässä K/N jollakin estimaatillaan. (Englannin kielessä tästä käytetään termiä *plug-in estimate*; sopiva suomenkielinen vastine olisi ehkä *sijoitus-* tai *kytkentäestimaatti*.) Erityisesti, jos sijoitamme K:n paikalle sen havaintoa $\boldsymbol{x}_n$ vastaavan suurimman uskottavuuden estimaatin, vastaus saa muodon "todennäköisyys sille, että seuraavalla nostolla saadaan valkoinen pallo, on $\hat{K}(\boldsymbol{x}_n)$".

On kuitenkin huomattava, ettei tällaista luonnehdintaa voi perustella täsmällisesti mallin $\boldsymbol{M}$ puitteissa — ellemme sitten päätä ohittaa koko ongelmaa tuntemattomasta parametrin arvosta ja valita siksi mallissa suoraan $k = \hat{K}(\boldsymbol{x}_n)$, jolloin voimme myös siirtyä käyttämään vastaavaa todennäköisyyttä P_k. Konkreettisessa koetilanteessa emme kuitenkaan voi tietää, vastaako tämä valinta kulhossa olevien valkoisten pallojen todellista lukumäärää. Tässä mielessä kytkentäestimaatin käyttöön liittyy aina ylimääräinen epävarmuustekijä, jota mallin $\boldsymbol{M}$ puitteissa ei ole voitu ottaa huomioon todennäköisyytenä.

Tarkastellaan sitten vastaava ennustamistehtävää mallin $\boldsymbol{M}^*$ puitteissa, jolloin myös muuttujan K saamiin arvoihin voidaan liittää todennäköisyyksiä. Ajatellaan aluksi tilannetta, jossa ei vielä ole tehty aiempia havaintoja: "Seuraava nosto" on silloin tietenkin järjestyksessä ensimmäinen. Saamme helposti (kokonaistodennäköisyyden kaavaa käyttäen, vrt. aiemmin esillä ollut kaava (1.5)) tuloksen

$$P(X_1 = 1) = \sum_k P(K = k)P(X_1 = 1 \mid K = k) = \frac{1}{N}\sum_k k\, P(K = k) = \frac{E(K)}{N}, \quad (1.10)$$

missä $E(K)$ on muuttujan K odotusarvo mallissa $\boldsymbol{M}^*$. Vastaavasti $P(X_1 = 0) = 1-\frac{E(K)}{N}$. Näin siis mallin $\boldsymbol{M}$ mukaan määritetyt valkoisen ja mustan pallon todennäköisyydet K/N ja $1-K/N$, missä K on mallin $\boldsymbol{M}$ parametri, tulevat mallissa $\boldsymbol{M}^*$ korvatuiksi lausekkeilla, joissa tuntemattoman K:n tilalla on sen odotusarvo.

Huomautus 1.3 *Tämä tulos on itse asiassa erikoistapaus* ($n = 1$) *reunatodennäköisyydestä, joka saadaan kaavasta* (1.3)*) summaamalla se yli kaikkien muuttujan K saamien arvojen, ts. todennäköisyydestä*

$$P(\boldsymbol{X}_n = \boldsymbol{x}_n) = \sum_{k=0}^{N} P(K = k)\, P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k). \quad (1.11)$$

Huomaa, että sama todennäköisyys esiintyi jo kaavan (1.8) *oikean puolen nimittäjässä ja että ehdolliset todennäköisyydet $P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k)$ saadaan yksinkertaisessa tulomuodossa lausekkeesta* (1.4)*. Kaavan* (1.11) *oikea puoli voitaisiin kirjoittaa myös lyhyesti muotoon $E(P(\boldsymbol{X}_n = \boldsymbol{x}_n|K))$, missä merkintä $P(\boldsymbol{X}_n = \boldsymbol{x}_n|K)$ tarkoittaa satunnaismuuttujaa, joka saa arvon $P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid K = k)$ silloin kun $K = k$. Tämä muuttuja on siis mallissa $\boldsymbol{M}^*$ satunnaismuuttujaksi tulkitun K:n funktio, ja odotusarvo "E" lasketaan K:n (priori)jakauman avulla samalla tavalla kuin edellä kaavan* (1.10) *oikealla puolella.*

Tutkitaan sitten, miten tilanne muuttuu yleisessä tapauksessa, kun käytettävissä ovat kokeessa tehdyt aiemmat havainnot $\boldsymbol{X}_n = \boldsymbol{x}_n$ ja "seuraava havainto" on X_{n+1}. Tarkastelemme siis *ennuste- eli prediktiivistä todennäköisyyttä*

$$P(X_{n+1} = x_{n+1} \mid \boldsymbol{X}_n = \boldsymbol{x}_n). \quad (1.12)$$

Aluksi on syytä palauttaa mieliin sivulla 9 tehty huomautus, etteivät muuttujat X_i ole riippumattomia mallissa $\boldsymbol{M}^*$. Erityisesti *ei siis päde yleisesti*, että

$$P(X_{n+1} = x_{n+1} \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = P(X_{n+1} = x_{n+1}).$$

Todennäköisyyden (1.12) määrittämiseksi on seuraavat kaksi reittiä (vaikkakin lopputulos on luonnollisesti sama riippumatta siitä, kumpaa kuljetaan): Voimme joko

(i) kirjoittaa

$$P(X_{n+1} = x_{n+1} \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = \frac{P(\boldsymbol{X}_n = \boldsymbol{x}_n, X_{n+1} = x_{n+1})}{P(\boldsymbol{X}_n = \boldsymbol{x}_n)} = \frac{P(\boldsymbol{X}_{n+1} = \boldsymbol{x}_{n+1})}{P(\boldsymbol{X}_n = \boldsymbol{x}_n)}$$

sekä käyttää sitten oikealla puolella osoittajassa kaavaa (1.11) kun siinä n:n tilalle kirjoitetaan $n + 1$, tai vaihtoehtoisesti

(ii) kirjoittaa jälleen kokonaistodennäköisyyden kaavaa käyttäen (nyt vain ehdollisille todennäköisyyksille $P(\cdot \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$ muotoiltuna)

$$\begin{aligned} &P(X_{n+1} = x_{n+1} \mid \boldsymbol{X}_n = \boldsymbol{x}_n) \\ &\quad = \sum_k P(K = k \mid \boldsymbol{X}_n = \boldsymbol{x}_n)\, P(X_{n+1} = x_{n+1} \mid K = k, \boldsymbol{X}_n = \boldsymbol{x}_n) \\ &\quad = \sum_k P(K = k \mid \boldsymbol{X}_n = \boldsymbol{x}_n)\, P(X_{n+1} = x_{n+1} \mid K = k), \end{aligned}$$

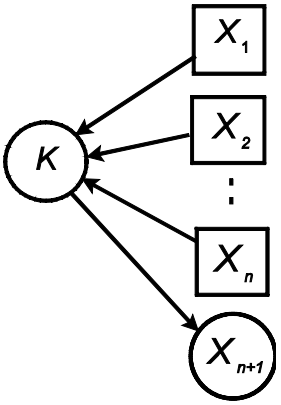
jolloin oikealla puolella olevassa summassa tulotermin ensimmäisenä tekijänä on edellä kaavalla (1.8) määritelty ja havaintoihin $\boldsymbol{X}_n = \boldsymbol{x}_n$ perustuva posterioritodennäköisyys tapahtumalle $K = k$. Toinen tekijä taas saa (vrt. kaava (1.1)) arvon k/N jos $x_{n+1} = 1$ ja arvon $1 - k/N$ jos $x_{n+1} = 0$. Näin ollen tulos voidaan esittää yksinkertaisessa muodossa

$$\begin{aligned} P(X_{n+1} = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}_n) &= \frac{E(K \mid \boldsymbol{X}_n = \boldsymbol{x}_n)}{N} \\ P(X_{n+1} = 0 \mid \boldsymbol{X}_n = \boldsymbol{x}_n) &= 1 - \frac{E(K \mid \boldsymbol{X}_n = \boldsymbol{x}_n)}{N}. \end{aligned} \tag{1.13}$$

Tämä vastaa täysin aiempaa tulosta (1.10), mutta K:n odotusarvot on nyt laskettava ehdollisina, ts. posteriorijakauman suhteen. Huomaa myös, että (i)-vaihtoehdon mukaan meneteltäessä osamäärän $P(\boldsymbol{X}_{n+1} = \boldsymbol{x}_{n+1})/P(\boldsymbol{X}_n = \boldsymbol{x}_n)$ nimittäjä on juuri se "ehtotapahtuman todennäköisyys", joka jätetään kirjoittamatta näkyviin, kun posterioritodennäköisyydet $P(K = k \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$ esitetään kaavan (1.9) mukaisesti verrannollisuusmerkintää "$\propto$" käyttäen.)

Näiden kaavojen merkitys on siinä, että ne esimerkissä kuvatussa yksinkertaisessa toistokokeessa antavat täsmällisen ja todennäköisyyslaskennan kannalta eräässä mielessä normatiivisen muodon sille, miten muuttujaa K koskevia arvioita — ja vastaavasti seuraavaa koetulosta koskevia ennusteita — tulee päivittää induktiivisesti toistokokeesta aiemmin tehtyjen havaintojen perusteella.

Päättelyä voidaan jälleen havainnollistaa kuvalla 1.4. Tämä kuva on eräänlainen yhdistelmä kuvista 1.2 ja 1.3. Muuttujien (havaintojen) X_i suhde muuttujaan K vastaa kuvan 1.3 mukaista estimointitehtävää, ts. päättelyn suunta (ja kuvassa sitä vastaavat nuolet) kulkevat muuttujista X_i muuttujaan K. Sitä vastoin muuttujien K ja X_{n+1} välinen suhde vastaa kuvan 1.2 mukaista tilannetta, jossa "X_{n+1}:n arvo poimitaan K:n määrittelemästä jakaumasta". Myös ennustetehtävä voidaan ajatella tällä tavalla jaetuksi kahteen vaiheeseen, joista ensimmäinen koskee parametriestimointia ja sitten toinen itse varsinaista ennustamista.



Kuva 1.4: Seuraavan havainnon X_{n+1} ennustaminen havaintojen $\boldsymbol{X}_n$ perusteella.

1.6 Jatkuvan parametrin tapaus: nasta lasipurkissa

Edellä käsitellyn johdattelevan esimerkin tarkastelut voidaan yleistää eri tavoin, jolloin ne varmaankin vastaavat paremmin käytännössä esiin tulevia erilaisia koe- ja oppimistilanteita. Jatkamme kuitenkin toisella hyvin yksinkertaisella konkreettisella esimerkillä, jota vastaava koe on helposti järjestettävissä luentotilanteessa.

Esimerkki 1.4 *Nastaa helistetään lasipurkissa ja sen jälkeen rekisteröidään, laskeutuuko se "selälleen" vai "kyljelleen". Tämän jälkeen koe toistetaan yhä uudelleen.*

Kappaleen 1.1 esimerkin antamaa mallia noudattaen merkitsemme nyt $X_i = 1$, jos nasta i:nnellä helistyksellä tulee selälleen, ja $X_i = 0$, jos se tulee kyljelleen. Jonkun verran ongelmallisempi sen sijaan on kysymys sopivan *parametrin* valinnasta. Nastaan lasipurkissa ei näyttäisi liittyvän mitään samalla tavalla ilmeistä parametria, jonka arvon voisimme määrittää suorittamalla esim. jokin sopiva mittaus, saati sitten että tällainen määritys voitaisiin tehdä vain purkkiin katsomalla.

Yksi mieleen tuleva mahdollisuus olisi ajatella, että kappaleen 1.1 nosta pallo kulhosta -esimerkissä valkoisten pallojen suhteellinen osuus, K/N, tulkittaisiin eräänlaiseksi palloja sisältävän kulhon taipumukseksi tuottaa nostettaessa valkoinen pallo. Vastaavasti voitaisiin ajatella, että nastalla on helistyksen päätyttyä tietty taipumus laskeutua selälleen. Tällä todennäköisyyden ns. *propensiteettitulkinnalla* pyritään jollakin tavalla "objektivisoimaan" todennäköisyyden käsite sellaisissakin tapauksissa, joissa ei voida järkevästi ottaen vedota fysikaaliseen symmetriaan ja joihin sen vuoksi klassinen todennäköisyyden käsite on soveltumaton. (Esillä olevassa tapauksessa ei esimerkiksi ole mitään perustetta olettaa, että tapahtumat "nasta laskeutuu selälleen" ja "nasta laskeutuu kyljelleen" olisivat yhtä todennäköiset. Toisaalta voi olla vaikeaa päätellä etukäteen, ilman varsinaisia havaintoja, kumpi näistä tapahtumista olisi todennäköisempi.)

Toinen, erityisesti toistokokeisiin soveltuva todennäköisyyden tulkinta nojaa säännellyissä koeolosuhteissa suoritetuissa pitkissä koesarjoissa havaittavaan suhteellisten frekvenssien stabiilisuuteen, joka on eräänlainen empiirinen vastine suurten lukujen laille. Tällä todennäköisyyden ns. *frekvenssitulkinnalla* on paljon kannatusta ja sen voidaan sanoa vastaavan useimpien tilastotieteilijöiden käsitystä järkevästä todennäköisyyskäsitteen tulkinnasta.

Oli lähtökohta näistä kahdesta kumpi tahansa, myös "nasta lasipurkissa" -esimerkin tapauksessa vaikuttaisi luontevalta valita mallin parametriksi välillä $(0,1)$ oleva reaaliluku θ, joka siten korvaisi aiemman esimerkin valkoisten pallojen suhteellisen osuuden K/N, mutta joka nyt vastaisi edellä mainittujen tulkintojen mukaisesti joko "nastan taipumusta laskeutua selälleen" tai "tällaisten koetulosten suhteellista osuutta äärettömän pitkissä koesarjoissa". Päädymme näin parametriseen tilastolliseen malliin

$$\boldsymbol{M} = \{P_\theta; 0 < \theta < 1\}.$$

Kullakin valitulla kiinteällä θ:n arvolla oletamme sitten muuttujat X_i riippumattomiksi todennäköisyyden P_θ suhteen, jolloin (aiempaa kaavaa (1.1) täysin vastaavalla päättelyllä) johdumme määrittelyyn

$$P_\theta(\boldsymbol{X}_n = \boldsymbol{x}_n) = \theta^{T(\boldsymbol{x}_n)}(1-\theta)^{n-T(\boldsymbol{x}_n)}, \qquad 0 < \theta < 1; \tag{1.14}$$

tässä on jälleen merkitty $T(\boldsymbol{x}_n) = \sum_{i=1}^n x_i$.

Muodollisesti kaikki se, mitä edellä käsitellyssä johdattelevassa esimerkissä todettiin mallin $\boldsymbol{M}$ suhteen, pätee myös tässä uudessa tilanteessa, kunhan vain osamäärä K/N korvataan parametrilla θ. Tämä pätee myös aiemmin esitetyn kuvan 1.1 suhteen.

Erityisesti voimme sanoa, että todennäköisyydet määrittelevä lauseke (1.14), kun se tulkitaan parametrin θ funktioksi, on mallin $\boldsymbol{M}$ mukainen havaintoa $\boldsymbol{x}_n$ vastaava *uskottavuusfunktio*. Samoin nähdään, että tämän uskottavuusfunktion maksimoiva parametrin arvo (ts. θ:n *suurimman uskottavuuden estimaatti*) saadaan kaavasta

$$\hat{\theta} = \hat{\theta}(\boldsymbol{x}_n) = \arg\max_\theta P_\theta\left(\boldsymbol{X}_n = \boldsymbol{x}_n\right) = \frac{1}{n}\sum_{i=1}^n x_i. \tag{1.15}$$

Koska $E_\theta(X_i) = P_\theta(X_i = 1) = \theta$, nähdään välittömästi, että tämä estimaattori on harhaton, ts. $E_\theta(\hat{\theta}(\boldsymbol{X}_n)) = \theta$, $0 < \theta < 1$. Vastaavasti nähdään helposti, että $\mathrm{Var}_\theta(\hat{\theta}(\boldsymbol{X}_n)) = \theta(1-\theta)/n$ (vrt. harjoitustehtävät 1.3 ja 1.4).

Tarkastellaan sitten edelleen samaa esimerkkiä 1.4, mutta tutkien mahdollisuutta, jossa myös θ voitaisiin tulkita satunnaismuuttujaksi. Kuten aiemmin parametrin K tapauksessa, luontevin tapa tehdä tämä on käyttää todennäköisyyden subjektiivista tulkintaa, jolloin θ:n arvoon liittyvän satunnaisuuden ajatellaan vastaavan "minun epävarmuuttani oikeasta parametrin arvosta". Kysymys "oikeasta parametrin arvosta" on puolestaan luontevaa kytkeä (hypoteettisissa) äärettömän pitkissä koesarjoissa saataviin "ykkösten" suhteelliseen frekvenssin.

Ennen kuin tätä ajatusta vastaava mallimäärittely voidaan tehdä täsmällisesti, on huomattava, että nyt malliparametri θ, satunnaismuuttujaksi tulkittuna, on *jatkuva* (eikä *diskreetti* kuten K). Tästä syystä mm. kaavassa (1.11) esiintyvät prioritodennäköisyydet $P(K = k)$ täytyy korvata vastaavilla välillä $(0,1)$ määritellyillä *todennäköisyystiheyksillä*. Käytämme seuraavassa näille tiheysfunktioille merkintöjä $p(\theta)$. Jos nyt tarkastelemme satunnaismuuttujaparia $(\theta, \boldsymbol{X}_n)$, on sen ensimmäinen koordinaatti θ jatkuva ja toinen koordinaatti $\boldsymbol{X}_n$ diskreetti satunnaismuuttuja. Näiden muuttujien yhteisjakauma, ja sen mukainen tilastollinen malli $\boldsymbol{M}^*$, voidaan kuitenkin edelleen määritellä todennäköisyyslaskennan kertolaskusääntöä käyttäen, kunhan vain muistetaan, että jatkuvien muuttujien — tässä siis θ:n — kohdalla todennäköisyyksien yhteenlasku tapahtuu laskemalla vastaavan määrätyn integraalin arvo. Täten siis voimme laskea esim. mallin $\boldsymbol{M}^*$ mukaiset havaintoihin $\boldsymbol{X}_n = \boldsymbol{x}_n$ liittyvät reunatodennäköisyydet integraalina

$$P(\boldsymbol{X}_n = \boldsymbol{x}_n) = \int_0^1 p(\theta)\, P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \theta)\, d\theta = \int_0^1 p(\theta)\, \theta^{T(\boldsymbol{x}_n)}\, (1-\theta)^{1-T(\boldsymbol{x}_n)}\, d\theta, \tag{1.16}$$

missä olemme määritelleet, että mallin $\boldsymbol{M}^*$ ehdollinen todennäköisyys $P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \theta)$ on sama kuin edellä mallin $\boldsymbol{M}$ puitteissa määritelty todennäköisyys (1.14). (Huom. Tässä ei ole θ:n kohdalla merkinnällä erotettu "satunnaismuuttujaa" ja "satunnaismuuttujan saamaa arvoa". Ehdollisen todennäköisyyden $P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \theta)$ on kuitenkin ymmärrettävä vastaavan ehtoa "satunnaismuuttuja θ saa arvon θ".)

Myös aiemmin kaavoissa (1.8) ja (1.9) vasemmalla puolella esiintyvät posterioritodennäköisyydet $P(K = k \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$ täytyy nyt korvata vastaavilla välillä $(0, 1)$ määritellyillä todennäköisyystiheydellä $p(\theta|\boldsymbol{X}_n = \boldsymbol{x}_n)$, jota voidaan kutsua *posteriori-tiheysfunktioksi*. Siten kaava (1.9) korvautuu nyt kaavalla

$$p(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n) \propto p(\theta)\, P(\boldsymbol{X}_n = \boldsymbol{x}_n \mid \theta) = p(\theta)\, \theta^{T(\boldsymbol{x}_n)}(1-\theta)^{n-T(\boldsymbol{x}_n)}, \qquad 0 < \theta < 1. \tag{1.17}$$

Tässä merkitsemättä jätetty θ:sta riippumaton tekijä $P(\boldsymbol{X}_n = \boldsymbol{x}_n)$ on juuri edellä määritetty reunatodennäköisyys (1.16). Vastaavasti edellä saadut ennustetodennäköisyydet korvautuvat nyt todennäköisyyksillä

$$P(X_{n+1} = x_{n+1} \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = \int_0^1 p(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n)\, \theta^{x_{n+1}}(1-\theta)^{1-x_{n+1}}\, d\theta. \tag{1.18}$$

Tapauksessa $x_{n+1} = 1$ saamme tuloksen

$$P(X_{n+1} = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = E(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$$

ja vastaavasti tapauksessa $x_{n+1} = 0$ tuloksen

$$P(X_{n+1} = 0 \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = 1 - E(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n),$$

missä odotusarvot määritetään posteriorijakauman (1.17) perusteella (vrt. kaava (1.13), missä θ:n tilalla on nyt K/N).

Pohdintaa. Todennäköisyyden $P(\boldsymbol{X}_n = \boldsymbol{x}_n)$ esitysmuotoon integraalina (1.16) voitaisiin päätyä myös toista kautta, käyttäen ns. vaihdettavuuden käsitettä. Muuttujajonoa $X_1, X_2, \ldots, X_n, \ldots$ sanotaan (*äärettömästi*) *vaihdettavaksi* (engl. *exchangeable*), jos kaikilla n:n arvoilla (vektori)muuttujan $(X_1, X_2, \ldots, X_n)$ todennäköisyysjakauma säilyy muuttumattomana vaikka sen koordinaattien $X_1, X_2, \ldots, X_n$ järjestystä vaihdettaisiin mielivaltaisella tavalla. Esimerkiksi edellä käsitellyn "nastan helistyksen" tapauksessa tällainen oletus vaikuttaa hyvin luonnolliselta: Mikäli purkkia aina helistetään perusteellisesti, osatulosten keskinäisen järjestyksen ei pitäisi järkevästi ottaen vaikuttaa todennäköisyyteen, joka liitetään mielivaltaisesti valittua pituutta n olevaan tulospöytäkirjaan. Ns. *deFinettin esityslause* (sovellettuna binäärimuuttujien tapaukseen) sisältää nyt tuloksen, jonka mukaan vaihdettavuushypoteesin pätiessä todennäköisyydet $P(\boldsymbol{X}_n = \boldsymbol{x}_n)$ voidaan esittää kaavan (1.16) mukaisella tavalla integraaleina. Huomaa, että tässä tuloksessa (jonka todistus tällä kurssilla sivuutetaan) päättelyn suunta ei ole "annetuista priorijakaumasta ja uskottavuusfunktiosta todennäköisyyksiin" vaan päinvastoin "(subjektiivisesta) todennäköisyydestä P priorijakaumaan $p(\theta)$ ja uskottavuusfunktioon (1.14)". On myös tärkeää huomata, että parametri θ on kaavoissa (1.16) ja (1.18) vain integroimismuuttuja. Sillä ei siten tämän lähestymistavan puitteissa edellytetä olevan mitään välitöntä tulkintaa esimerkiksi "yhteen koetulokseen" liittyvän tapahtuman $\{X_n = 1\}$ (propensiteetti)todennäköisyytenä. Itse asiassa kaavasta (1.16) nähdään, että

$$P(X_1 = x_1) = \int_0^1 p(\theta)\, \theta^{x_1}(1-\theta)^{1-x_1}\, d\theta, \tag{1.19}$$

mikä vastaa tulosta (1.10) aiemman esimerkkimme yhteydessä. Koetuloksen $X_1 = 1$ todennäköisyys mallissa $\boldsymbol{M}^*$ on siis

$$P(X_1 = 1) = \int_0^1 \theta\, p(\theta)\, d\theta,$$

eli (satunnaismuuttujaksi tulkitun) θ:n odotusarvo priorijakauman suhteen. Vaihdettavuuden perusteella sama reunatodennäköisyys pätee luonnollisesti kaikille muillekin muuttujille X_n yksinään tarkasteltuina, ellei varsinaisia havaintoja ole käytettävissä. Voidaan kuitenkin osoittaa, että vaihdettavuusoletuksen vallitessa koetuloksen "1" suhteelliset frekvenssit $\frac{1}{n}T(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^n X_i$ konvergoivat n:n kasvaessa (P-todennäköisyydellä 1). Jos tätä raja-arvoa merkitään θ:lla, voidaan esityslauseessa (1.16) oleva tiheysfunktio $p(\theta)$ luonnollisella tavalla tulkita henkilön ennakkokäsitykseksi θ:n arvosta, jossa ilmenevä θ:n satunnaisuus voidaan sitten ymmärtää epävarmuudeksi siitä, mitä raja-arvoa kohti konvergenssi tapahtuu. Tämä deFinettin esityslauseeseen pohjautuva tarkastelu muodostaa — vaihdettavuusoletuksen vallitessa — eräänlaisen sillan todennäköisyyden subjektiivisen ja frekventistisen tulkinnan välille: Jos edellä kaavassa (1.19) prioritodennäköisyys(tiheys) $p(\theta)$ korvattaisiin jossakin tietyssä pisteessä θ_0 olevalla pistetodennäköisyydellä, tämän kaavan oikea puoli muuntuisi muotoon $\theta_0^{x_1}(1-\theta_0)^{1-x_1}$, joka on ennestään tuttu binomitodennäköisyys parametrien arvoilla $n = 1$ ja θ_0. Vastaava yleisempi n:n havainnon pöytäkirjaa koskeva muoto saadaan kaavasta (1.16).

Käytännön kannalta on tietenkin mukavaa, jos edellä esitetyt integraalit voidaan laskea jollakin analyyttisesti yksinkertaisella tavalla. Jos uskottavuusfunktio on edellä esitettyä muotoa (1.14), on kätevää valita priorijakauma ns. *Beta*-jakaumaperheestä, jonka tiheysfunktiot ovat muotoa

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \text{missä } 0 < \theta < 1,$$

ja $\alpha > 0$ ja $\beta > 0$ ovat jakauman parametreja. Tässä merkitsemättä jäänyt normeeraustekijä — koska $p(\theta)$ on tiheysfunktio — on luonnollisesti $(B(\alpha,\beta))^{-1}$, missä on merkitty

$$B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta.$$

Tätä integraalia kutsutaan nimellä (Eulerin) beeta-funktio, ja sen arvot eri α:n ja β:n arvoilla löytyvät monista taulukoista. Numeeristen arvojen laskemiseksi on hyödyllistä käyttää tietoa, jonka mukaan $B(\alpha,\beta)$ voidaan esittää muodossa $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$; tässä ns. (Eulerin) gamma-funktio toteuttaa ehdot

$$\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1) \qquad \text{kaikilla } \alpha > 1$$

ja $\Gamma(1) = 1$, joten α:n kokonaislukuarvoilla $\Gamma(\alpha) = (\alpha-1)!$. Tapauksessa $\alpha = \beta = 1$ saadaan välillä $(0,1)$ määritelty *tasajakauma* $p(\theta) = 1$, $0 < \theta < 1$.

Verrannollisuusmerkintää "$\propto$" käyttäen nähdään välittömästi, että kun priorijakauma valitaan Beta-jakaumaperheestä, myös posteriorijakauma on Beta-jakauma, mutta parametrien arvot α ja β ovat nyt muuttuneet arvoiksi $\alpha + \sum_{i=1}^n x_i$ ja $\beta + (n - \sum_{i=1}^n x_i)$. Tällaista priorijakuman ja uskottavuusfunktion välistä suhdetta, jossa posteriorijakauma säilyy samassa jakaumaperheessä kuin mihin priorijakaumakin kuuluu, luonnehditaan sanomalla, että ne ovat toistensa *liittojakaumia* (engl. *conjugate distributions*).

Edellä olevat "nasta lasipurkissa" -esimerkkiin liittyvät tarkastelut voidaan nyt välittömästi yleistää tilanteisiin, joita voi luonnehtia seuraavasti:

- kyseessä on toistokoe, jossa tiettyä koetta toistetaan samanlaisissa olosuhteissa jokin äärellinen määrä kertoja
- kussakin kokeessa erotetaan kaksi tulosvaihtoehtoa
- peräkkäisten toistokokeiden tulokset oletetaan joko toistaan riippumattomiksi, kun koetta kuvaava parametrinarvo on annettu (malli $\boldsymbol{M}$), tai yleisemmin vaihdettaviksi (malli $\boldsymbol{M}^*$).

Tällaista koetta kutsutaan *binomikokeeksi*.

Luku 2

Jakaumaperheiden tilastotiedettä

Varsin tavallista on, että tilastollisen päättelyn lähtökohdaksi valitaan jokin ns. *parametrinen tilastollinen malli*. Edellä käsitellyt esimerkit "palloja kulhossa" ja "nasta lasipurkissa" johtivat kumpikin periaatteessa samanlaiseen *binomikoe*-malliin. Todennäköisyyslaskennan kannalta mallien olennaisin ero oli, että jälkimmäisessä parametri θ oli *jatkuva* muuttuja, saaden kaikkia reaaliarvoja välillä $[0, 1]$, kun taas edellisessä parametri K oli kokonaislukuarvoinen ja siten *diskreetti* muuttuja. (Myös tässä tapauksessa parametri voidaan luonnollisesti skaalata välille $[0, 1]$ siirtymällä käyttämään parametrina valkoisten pallojen suhteellista osuutta K/N.) Tilastollisen päättelyn kannalta olennaisempi ero näiden esimerkkien välillä on kuitenkin niiden parametrien tulkinnassa: parametrilla K on ainakin periaatteessa suoran havainnon kautta yksikäsitteisesti määräytyvä arvo, kun taas parametrin θ kohdalla tilanne on paljon monisyisempi ja käsitteellisesti vaikeammin jäsennettävissä.

Ääritapauksessa, jos lähtökohdaksi otetaan deFinettin esittämä henkilökohtainen todennäköisyys ja samalla oletetaan koetoistojen antamien tulosten vaihdettavuus tämän todennäköisyyden suhteen, yksittäisille parametrinarvoille ei välttämättä tarvitse antaa mitään konkreettista fysikaalista tulkintaa. Parametri θ voidaan silloin ymmärtää pelkkänä integroimismuuttujana, joka kaikissa havaittavia muuttujia koskevissa todennäköisyystarkasteluissa, kuten edellä lausekkeissa (1.16) ja (1.18) "integroidaan pois".

Ajatusprosessi, jonka tuloksena tutkija päätyy kuvaamaan tarkastelemaansa ilmiötä jonkin tietyn tilastollisen mallin avulla, on useimmiten monivaiheinen eikä siihen voi liittää mitään selkeitä päättelysääntöjä tai ohjeita. Tavallisesti käyttökelpoiset mallit kuitenkin noudattavat eräänlaista modulirakennetta, jonka osat sitten valitaan analyyttisesti suhteellisen helposti hallittavista ja parametrisesti määritellyistä jakaumaperheistä, pitäen samalla silmällä havaittavien satunnaismuuttujien käyttäytymistä, niiden arvoalueita, ym.

Tarkastelemme nyt eräitä tällaisia jakaumaperheitä (ks. myös luentokurssin *Johdatus todennäköisyyslaskentaan* muistiinpanot).

2.1 Tärkeitä diskreettejä jakaumia

Binomijakauma (merk. "Bin(n, θ)") syntyy tarkastelemalla ykkösten lukumäärää em. *binomikoe*-tilanteessa, kun koetta toistetaan jokin kiinteä määrä kertoja. Muodollisesti binomimuuttuja voidaan määritellä summana $T(\boldsymbol{X}_n) = \sum_{i=1}^{n} X_i$, missä muuttujat X_i ovat mallissa $\boldsymbol{M}$ riippumattomia $\{0, 1\}$-arvoisia satunnaismuuttujia todennäköisyyksien P_θ suhteen. Näin binomijakauman parametriksi on luontevaa valita lukupari (n, θ). Useimmissa käytännön tilanteissa toistojen lukumäärä n valitaan etukäteen, joten varsinai-

seksi malliparametriksi jää silloin θ. Jakaumaa vastaavat pistetodennäköisyydet $p_\theta(x)$, $x = 0, 1, \ldots, n$, määritellään tunnetulla tavalla kaavalla

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \tag{2.1}$$

missä *binomikerroin* $\binom{n}{x} = \frac{n!}{(n-x)!x!}$ on niiden nollista ja ykkösistä muodostuvien havaintopöytäkirjojen $(x_1, x_2, \ldots, x_n)$ lukumäärä, joille $\sum_{i=1}^n x_i = x$.

Huomautus 2.1 *Joskus on tarkoituksenmukaisempaa merkitä jakauman parametri esimerkiksi puolipisteen jälkeen sulkumerkkien sisään, varaten sitten alaindeksi viittaamaan siihen satunnaismuuttujaan, jonka jakaumasta kulloinkin on kysymys. Tätä merkintätapaa noudattaen voitaisiin edellä kirjoittaa kaavan* (2.1) *vasemmalle puolelle* $p_X(x;\theta)$. *Vastaavat odotusarvon ja varianssin lausekkeet ovat binomijakauman tapauksessa*

$$E(X;\theta) = n\theta \quad ja \quad \mathrm{Var}(X;\theta) = n\theta(1-\theta).$$

Binomimuuttujan edellä annetusta määritelmästä summana $\sum_{i=1}^n X_i$ näkyy suoraan seuraava niiden *additiivisuusominaisuus*: Olkoon $X_1 \sim \mathrm{Bin}(n_1, \theta)$ ja $X_2 \sim \mathrm{Bin}(n_2, \theta)$, ja oletetaan ne riippumattomiksi mallissa $\boldsymbol{M}$. Silloin niiden summa $Y = X_1 + X_2$ on $\mathrm{Bin}(n_1 + n_2, \theta)$-jakautunut.

Poisson-jakautunut satunnaismuuttuja X saa kokonaislukuarvoja $0, 1, 2, \ldots$. Siihen liitettävät pistetodennäköisyydet määritellään kaavalla

$$p(x;\mu) = e^{-\mu} \frac{\mu^x}{x!}, \tag{2.2}$$

missä parametri $\mu > 0$ on samalla muuttujan X odotusarvo, ts. $E(X;\mu) = \mu$. Suoralla laskulla nähdään, että myös $\mathrm{Var}(X;\mu) = \mu$. Poisson-jakauma (merk. "Poisson(μ)") esitetään usein rajajakaumana binomijakaumasta $\mathrm{Bin}(n, \theta)$, kun jälkimmäisessä samanaikaisesti $n \to \infty$ ja $\theta \to 0$ siten, että niiden tulo $n\theta$ suppenee kohti jotakin vakioarvoa μ. Tästä vakioarvosta tulee sitten Poisson-jakauman parametri.

Binomijakauman tapaan myös Poisson-jakaumalla on seuraava additiivisuusominaisuus: Olkoon satunnaismuuttujilla X_1 ja X_2 jakaumat Poisson(μ_1) ja Poisson(μ_2). Oletetaan ne riippumattomiksi mallissa $\boldsymbol{M}$. Silloin niiden summa $Y = X_1 + X_2$ on Poisson($\mu_1 + \mu_2$)-jakautunut. Tämä tulos voidaan perustella usealla eri tavalla. Yksinkertainen perustelu voisi olla seuraava: Jos Poisson-muuttujat tulkitaan edellä kuvatulla tavalla Binomimuuttujien approksimaatioiksi, myös Binomi-muuttujien additiivisuusominaisuus periytyy approksimaatiossa Poisson-muuttujille. Täsmällisemmin samaan johtopäätökseen voi kuitenkin päätyä suhteellisen helposti myös suoralla laskulla pilkkomalla tapahtuma $\{Y = y\}$ aluksi muotoa $\{X_1 = x, X_2 = y - x\}$, $x = 0, 1, 2, \ldots, y$, oleviin osiin ja käyttämällä sitten oletusta muuttujien X_1 ja X_2 riippumattomuudesta sekä kokonaistodennäköisyyden kaavaa.

Poisson-jakaumaa käytetään useimmiten mallintamaan ns. *laskuriprosesseja*, ts. tilanteissa, joissa seurataan ajassa etenevänä prosessina satunnaisesti ja toisistaan riippumatta ilmaantuvia sykäyksiä, rekisteröiden samalla niiden kokonaismäärää. Klassisen esimerkin tällaisesta tilanteesta muodostavat jossakin tarkasteltavassa säteilylähteessä radioaktiivisen hajoamisen aiheuttamat impulssit ja niiden havainnointi Geiger-laskimen avulla. Tällaisissa tapauksissa on luontevaa esittää parametri muodossa $\mu = \lambda t$, missä λ on

säteilylähteen *intensiteetti* (jonka tässä oletetaan pysyvän ajan suhteen vakiona) ja t on seurantajakson pituus. Myös itse satunaismuuttujan merkinnässä voi olla paikallaan ottaa huomioon, kuinka pitkältä ajalta havaintoja on kerätty, kirjoittaen esim. $N(t)$ muuttujan X sijaan. Kun muuttujia $N(t)$ tarkastellaan realisaatioittain ajan t funktiona, päädytään tällä tavoin ns. *Poisson-prosessiin* $\{N(t); t > 0\}$. (Poisson-prosessia käsitellään laajemmin stokastisten prosessien luentokurssilla.)

Muita usein käytettyjä diskreettejä jakaumia ovat esim. *geometrinen* ja *hypergeometrinen* jakauma sekä *negatiivinen binomijakauma*.

2.2 Uskottavuusfunktion muodostaminen diskreettien jakaumien tapauksessa

Tilastollisen estimointiongelman eräänlaisena prototyyppinä voidaan pitää esim. binomikoetta vastaavan parametrin θ arviointia annetun havaintoaineiston perusteella, joka muodostuu riippumattomasti suoritetulla otannalla. Sanaa "riippumaton" käytetään tällöin usein kuvaamaan mekanismia tai tapaa, jolla otanta suoritetaan, mutta ilman että tälle termille olisi annettu selvää kytkentää tarkasteltuun tilastolliseen malliin. Todennäköisyyslaskennan kannalta on kuitenkin tärkeää kysyä, minkä todennäköisyyden suhteen tällainen otokseen sisältyvien havaintojen välinen riippumattomuusominaisuus ehkä pätee.

Perinteisessä *frekventistisessä* tilastollisessa päättelyssä aineiston sisältämien havaintojen $X_1, X_2, \ldots, X_n$ välinen riippumattomuusominaisuus oletetaan tavallisesti päteväksi suhteessa johonkin parametrisesti määriteltyyn todennäköisyyteen P_θ, jolloin voimme lyhyesti sanoa sen pätevän mallissa $\boldsymbol{M} = \{P_\theta; \theta \in \Theta\}$. Tällöin kutakin valittua parametrin θ arvoa vastaavan jakauman P_θ katsotaan kuvaavan jonkin tutkittavan ja — parhaassa tapauksessa — mitattavissa olevan ominaisuuden vaihtelua eri yksilöiden, kokeiden tai muiden havaintokohteiden välillä jossakin (ideaalitapauksessa kooltaan äärettömässä, mutta käytännössä "hyvin suuressa") *perusjoukossa* eli *populaatiossa*. Parametria θ sen sijaan pidetään tässä tulkinnassa populaation ominaisuutena ja parametrin θ vaihtelua arvojoukossa (*parametriavaruudessa*) Θ pidetään ilmauksena siitä, minkälaiset populaatiot ongelmassa tulevat periaatteessa kysymykseen. Teknisesti kysymys mallin määrittelystä ratkaistaan tavallisesti niin, että aluksi valitaan tarkasteltavan tilanteen kannalta tarkoituksenmukainen yhtä havaintoa vastaava pistetodennäköisyys $p_{X_i}(x_i; \theta)$ ja laajennetaan malli sitten luonnollisella tavalla vastaamaan koko havaintoaineistoa $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ vedoten tällöin havaintojen väliseen riippumattomuusominaisuuteen ao. mallissa. (Huom. Bayes-päättelyssä, jossa toimitaan periaatteessa yhden (tulkinnaltaan henkilökohtaisen) todennäköisyyden puitteissa, havaintojen välinen riippumattomuus — mikäli sellaiseen halutaan vedota — tulkitaan tämän todennäköisyyden suhteen ehdollisena, ts. "kun θ on annettu").

Diskreettien havaintomuuttujien ja riippumattoman otannan tapauksessa mallin $\boldsymbol{M}$ mukainen uskottavuusfunktio voidaan sitten laskea yksinkertaisesti tulona, jonka tekijät vastaavat kukin otoksen yhteen havaintoon $X_i = x_i$, $i = 1, 2, \ldots, n$, liitettävää uskottavuusfunktion arvoa. Voimme siis diskreettien jakaumien tapauksessa kirjoittaa riippumatonta otosta $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ vastaavan uskottavuusfunktion lausekkeen tulona

$$L(\theta) = P_\theta(\boldsymbol{X}_n = \boldsymbol{x}_n) = \prod_{i=1}^{n} P_\theta(X_i = x_i) = \prod_{i=1}^{n} p_{X_i}(x_i; \theta). \tag{2.3}$$

Suurimman uskottavuuden estimaattorin lauseke määräytyy nyt samalla tavalla kuin edellä kaavassa (1.15), ts.

$$\hat{\theta} = \hat{\theta}(\boldsymbol{x}_n) = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} P_\theta(\boldsymbol{X}_n = \boldsymbol{x}_n).$$

Huomautus 2.2 *Sellaisissa tapauksissa, joissa havaintojen väliseen riippumattomuusominaisuuteen mallissa $\boldsymbol{M}$ ei voi vedota ja tulomuoto* (2.3) *ei näin ollen tule kysymykseen, uskottavuusfunktion lauseke on periaatteessa aina laskettavissa käyttämällä todennäköisyyslaskennan kertolaskusääntöä ja kirjoittamalla (vrt. kaavaa* (1.1) *seuraava huomautus)*

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta) \\ &\quad = \prod_{i=1}^{n} P(X_i = x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}; \theta). \end{aligned} \tag{2.4}$$

Käytännössä voidaan lisäksi varsin usein vedota johonkin muuttujien X_i välillä vallitsevaan ehdolliseen riippumattomuusominaisuuteen, jolloin ainakin osa kaavan (2.4) *oikean puolen ehdollisista todennäköisyyksistä yksinkertaistuu. Erityisesti näin on laita ns.* Markovin ketjuissa*, joissa oletetaan ominaisuus*

$$P(X_i = x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}; \theta) = P(X_i = x_i \mid X_{i-1} = x_{i-1}; \theta). \tag{2.5}$$

Havainnollisesti sanoen: "Kun ennustetaan huomista, Markovin ketjun muisti ulottuu tähän päivään, mutta ei enää eiliseen". Markovin ketjuun liitettävät todennäköisyydet voidaan näin hallita määrittelemällä vain "yhtä päivää vastaavat" ns. siirtymätodennäköisyydet. (Markovin ketjuja tarkastellaan yksityiskohtaisemmin stokastisten prosessien luentokurssilla.)

Harjoitustehtävä 2.1 *Laske Beta(α, β)-jakauman odotusarvo ja varianssi. Luonnostele sen tiheysfunktion kuvaajat tarkastelemalla α:n ja β:n arvoja* 0.5, 1 *ja* 2. *[Huom. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.]*

Harjoitustehtävä 2.2 *Valmista itsellesi jokin havaintoväline, johon liittyviä peräkkäisiä havaintoja on luontevaa kuvata binomikokeena (esim. "nasta lasipurkissa", mutta omaa mielikuvitusta suositellaan käytettävän — tuo havaintovälineesi mukaan laskuharjoituksiin!). Kuvaa ennakkokäsitystäsi ao. binomikokeen malliparametrista θ käytämällä Beta(α, β)-jakaumaa, jossa valitset sopivat α:n ja β:n arvot. Suorita sitten peräkkäin kolme koetta ja määritä näitä koetuloksia vastaavat ennustetodennäköisyydet $P(X_{n+1} = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$, $n = 1, 2, 3$. (Voit tarvittaessa jatkaa toistokoetta pitempäänkin, ennustaen aina seuraavan tulosta.)*

Harjoitustehtävä 2.3 *Olkoon $X_1, X_2, \ldots, X_n$ riippumaton otos Poisson(μ)-jakautuneita satunnaismuuttujia mallissa $\boldsymbol{M}$. Johda parametrin μ suurimman uskottavuuden estimaattorin lauseke. [Opastus: Tarkastele jälleen uskottavuusfunktion logaritmin lauseketta ja määritä sen maksimikohta derivoimalla. Huomaa jälleen, että estimaattorin lauseke riippuu saaduista koetuloksista vain niiden summan kautta, ts. summaa voidaan pitää mallissa μ:n suhteen tyhjentävänä tunnuslukuna.]*

Harjoitustehtävä 2.4 *Näytä, että edellisessä tehtävässä johdettu estimaattori on harhaton, ts. (satunnaismuutujaksi tulkittuna) sen odotusarvo on μ. Laske myös sen varianssi ja tarkastele, mitä raja-arvoa se lähestyy kun otoskoko n kasvaa. [Huom. Jälkimmäisen ominaisuuden perusteella saatua estimaattoria voidaan kutsua tarkentuvaksi, vrt. harjoitustehtävä 1.4.]*

2.3 Eräitä jatkuvia jakaumia

Tutkitaan sitten, miten edellä olevia tarkasteluja joudutaan muokkaamaan tapauksessa, joissa vastemuuttuja on diskreetin sijasta *jatkuva* (saaden arvoja reaaliakselilla tai sen jollakin välillä). Tässä tapauksessa tilastollinen malli $\boldsymbol{M}$ määritellään tavallisesti ottamalla lähtökohdaksi yhtä havaintoa X_i vastaava tiheysfunktio $f(x_i;\theta)$, parametrin θ vaihdellessa arvoalueessa Θ.

Vastemuuttujan suhteen tärkeitä jatkuvia jakaumia ovat mm. (ks. myös luentokurssin *Johdatus todennäköisyyslaskentaan* luentomateriaali):

Tasajakauma. Olkoot a ja b reaalilukuja ja $a < b$. Silloin voimme määritellä tiheysfunktion $f_X(x;a,b)$ siten, että se saa vakioarvon $\frac{1}{b-a}$ välillä (a,b) ja arvon 0 tämän välin ulkopuolella. Vastaava kertymäfunktio $F_X(x;a,b)$, jonka parametrina on lukupari (a,b), saa arvon 0 kun $x \leq a$, arvon $\frac{x-a}{b-a}$ kun $a < x < b$, ja arvon 1 kun $x \geq b$. On helppoa todeta, että tiheys- ja kertymäfunktioiden välinen yhteys

$$F_X(x;a,b) = \int_{-\infty}^{x} f_X(y;a,b)\,dy, \qquad x \in \mathbb{R}, \tag{2.6}$$

pätee kaikkialla ja että $F'_X(x;a,b) = f_X(x;a,b)$ kaikissa muissa pisteissä paitsi pisteissä a ja b, joissa F_X:llä on kyllä vasemman ja oikeanpuoleiset derivaatat, mutta ne eivät ole samoja. Näissä pisteissähän edellä määriteltiin f_X:n arvoiksi $f_X(a) = f_X(b) = 0$. Toisaalta tiheysfunktion arvoksi näissä kahdessa pisteessä olisi yhtä hyvin voitu valita arvoiksi esim. $f_X(a) = f_X(b) = \frac{1}{b-a}$ ilman, että kaavan (2.6) pätevyys olisi tästä mitenkään muuttunut.

Esimerkki 2.3 *Edellä käsiteltiin lyhyesti Beta-jakaumaperhettä, jonka tiheysfunktiot, jos niitä sovelletaan vastemuuttujaan X ja jätetään normeeraustekijä $B(\alpha,\beta)^{-1}$ merkitsemättä, ovat muotoa*

$$f_X(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1},$$

missä $0 < x < 1$ ja $\alpha > 0$ ja $\beta > 0$ ovat jakauman parametreja. Tällöin myös mainittiin, että valinta $\alpha = \beta = 1$ johtaa välillä $(0,1)$ määriteltyyn tasajakaumaan.

Eksponenttijakaumaa käytetään useimmiten kuvaamaan elinikää ja tästä syystä sitä vastaavaa satunnaismuuttujaa merkitään useimmiten kirjaimella T. Eksponenttijakautunut satunnaismuuttuja saa kaikkia positiivisia reaaliarvoja ja sen tiheysfunktio voidaan määritellä kaavalla

$$f_T(t;\lambda) = \lambda \exp(-\lambda t), \qquad t > 0, \tag{2.7}$$

missä jakauman parametria on merkitty symbolilla λ. Itse jakaumasta käytetään usein lyhennettä $\mathrm{Exp}(\lambda)$. Suoralla integroinnilla voidaan helposti todeta, että tiheysfunktiota (2.7) vastaavalla kertymäfunktiolla on lauseke

$$F_T(t;\lambda) = 1 - \exp(-\lambda t), \qquad t > 0,$$

ja että $E(T;\lambda) = 1/\lambda$ ja $\mathrm{Var}(T;\lambda) = 1/\lambda^2$.

Gamma-jakauman tiheysfunktio määritellään kaavalla

$$f_X(x;\alpha,\beta) = \frac{\exp(-\beta x)\,\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}, \qquad x > 0, \tag{2.8}$$

missä $\alpha > 0$ ja $\beta > 0$ ovat jakauman parametreja ja $\Gamma(\alpha)$ on aiemmin esillä ollut Eulerin gammafunktio. Tapauksessa $\alpha = 1$, koska $\Gamma(1) = 1$, gammajakauma yhtyy eksponenttijakaumaan. Yleisemmin voidaan näyttää, että jos α on positiivinen kokonaisluku ja muuttujat $T_1, T_2, \ldots, T_\alpha$ ovat riippumattomia ja niistä kukin noudattaa Exp(β)-jakaumaa, silloin niiden summa $X = T_1 + T_2 + \cdots + T_\alpha$ noudattaa yhtälön (2.8) mukaista gammajakaumaa. Toisaalta valinta $\alpha = n/2$, missä n on positiivinen kokonaisluku, ja $\beta = \frac{1}{2}$ antaa tulokseksi tilastollisessa päättelyssä varsin tärkeän $\boldsymbol{\chi}^2(n)$-jakauman (lue: "khii-toiseen, vapausastelukuna n"). Tähän jakaumaan palaamme myöhemmin tarkemmin.

Poisson-jakauman sekä gamma- ja eksponenttijakauman välisiä mielenkiintoisia yhteyksiä tullaan käsittelemään tarkemmin stokastisten prosessien luentokurssilla.

Harjoitustehtävä 2.5 *(*) Olkoon X* Poisson(μ)*-jakautunut satunnaismuuttuja. Tulkitaan sitten myös parametri μ satunnaismuuttujaksi ja oletetaan, että sen arvoon liittyvää epävarmuutta (ennen kuin koetulos $X = x$ on käytettävissä) voidaan kuvata (priori)jakaumalla Gamma(α, β). Näytä, että tällöin myös tämän havainnon avulla saatava posteriorijakauma on muodoltaan Gamma-jakauma. Määritä myös sen parametrit. [Opastus: Noudata samaa periaatetta kuin edellä tarkasteltaessa binomikokeesta saatavia havaintoja ja malliparametriin liitettävää Beta-(priori)jakaumaa. Huomaa, että myös nyt posteriorijakauma säilyy samassa jakaumaperheessä, ts. Gamma- ja Poisson jakauma ovat toistensa liittojakaumia.]*

2.4 Normaalijakauma

Normaalijakauman (vrt. kurssin *Johdatus todennäköisyyslaskentaan* luentomateriaali) määritelmän lähtökohtana voidaan pitää ns. standardi-normaalimuuttujaa Z, jonka tiheysfunktio määritellään kaavalla

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-z^2/2\right), \qquad -\infty < z < \infty. \tag{2.9}$$

Tästä jakaumasta käytetään usein merkintää $N(0,1)$. Merkinnässä luvut 0 ja 1 viittaavat siihen, että näin jakautuneen muuttujan odotusarvo on $E(Z) = 0$ ja varianssi $\mathrm{Var}(Z) = 1$. Näistä odotusarvoa koskeva kohta seuraa suoraan tiheysfunktion φ symmetriasta origon suhteen ja jälkimmäinen on helppo osoittaa osittaisintegroinnilla. Tämän jakauman kertymäfunktiota on tapana merkitä kirjaimella Φ. Kertymäfunktion määrittävää integraalia $\Phi(x) = \int_{-\infty}^{x} \varphi(y)\,dy$ ei voida esittää suljetussa muodossa ja tästä syystä sen arvot eri pisteissä x joudutaan määrittämään numeerisen integroinnin keinoin. Kertymäfunktion arvot on taulukoitu käytännöllisesti katsoen kaikkien tilastotieteen kurssikirjojen liiteosaan. Ne löytyvät myös mm. verkko-osoitteesta
http://www.statsoft.com/textbook/distribution-tables/.

Tarkastellaan sitten muuttujan Z muotoa

$$X = \mu + \sigma Z \tag{2.10}$$

olevia lineaarimuunnoksia, missä μ on mielivaltainen reaaliluku ja $\sigma > 0$. Lukuparia $\theta = (\mu, \sigma^2)$ voidaan pitää muuttujan X jakauman parametrina. Selvyyden vuoksi vastaavaa todennäköisyysjakaumaa voidaan merkitä $P_{(\mu,\sigma^2)}$:llä. Odotusarvon ja varianssin ominaisuuksien perusteella voidaan heti todeta, että μ on X:n odotusarvo ja σ^2 sen varianssi, ts. $E(X) = \mu$ ja $\mathrm{Var}(X) = \sigma^2$. Toisaalta muuttujan X jakaumakin voidaan sitoa yksinkertaisella tavalla standardi-normaalimuuttujan Z jakaumaan: Kirjoittamalla

$$\begin{aligned} F_X(x;\mu,\sigma^2) &= P_{(\mu,\sigma^2)}(X \le x) = P_{(0,1)}\left(\mu + \sigma Z \le x\right) \\ &= P_{(0,1)}\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned} \tag{2.11}$$

havaitaan, että X:n kertymäfunktion arvo pisteessä x voidaan määrittää suoraan standardi-normaalijakauman kertymäfunktion arvona pisteessä $\frac{x-\mu}{\sigma}$. Sanomme nyt, että edellä kaavalla (2.10) määritelty muuttuja jakautuu normaalisti, odotusarvona μ ja varianssina σ^2. Tätä merkitään usein lyhyesti $X \sim N(\mu, \sigma^2)$. Vastaava tiheysfunktio saadaan tästä derivoimalla x:n suhteen, jolloin tulokseksi tulee

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.12}$$

Kääntäen voidaan helposti osoittaa, että jos satunnaismuuttuja jakautuu tiheysfunktion (2.12) osoittamalla tavalla, niin se voidaan esittää muodossa (2.10), missä $Z = \frac{X-\mu}{\sigma}$ on standardi-normaalimuuttuja. Tällöin sanotaan usein, että $N(\mu, \sigma^2)$-jakautunut muuttuja X on tällä muunnoksella "standardoitu" $N(0, 1)$-jakautuneeksi.

Määritelmän (2.10) perusteella nähdään heti, että normaalijakauma säilyy lineaarimuunnoksissa, ts. jos $X \sim N(\mu, \sigma^2)$, muotoa $Y = a + bX$ olevat muuttujat jakautuvat normaalisesti merkinnän $N(a + b\mu, (b\sigma)^2)$ mukaisesti.

Voidaan myös osoittaa, että normaalijakauma säilyy riippumattomien normaalimuuttujien yhteenlaskussa: Jos $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ ja X_1 ja X_2 ovat riippumattomia, niin $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Tässä odotusarvon ja varianssin yhteenlaskuominaisuudet on todettu jo aiemmin.

Induktiopäättelyllä riippumattomien normaalimuuttujien yhteenlaskua koskeva tulos voidaan luonnollisesti yleistää kahden muuttujan tapauksesta mille tahansa n:n riippumattoman normaalimuuttujan joukolle. Jos erityisesti oletamme, että muuttujat $X_1, X_2, \ldots, X_n$ noudattavat kaikki samaa normaalijakaumaa $N(\mu, \sigma^2)$ ja ovat riippumattomia, pätee edellisen perusteella $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$. Toisaalta, jos summan $\sum_{i=1}^n X_i$ asemesta tarkastellaan näiden muuttujien aritmeettista keskiarvoa $\bar{X}_n$, voidaan tietenkin yhtäpitävästi todeta, että

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2.13}$$

Tätä huomattavasti mielenkiintoisempi on kuitenkin tulos, jonka mukaan normaalijakauma voi myös "syntyä", ainakin likimäärin, laskettaessa yhteen riippumattomia satunnaismuuttujia, vaikka yhteenlaskettavat itse eivät olisikaan normaalijakautuneita. Tämä ns. *todennäköisyyslaskennan keskeinen raja-arvolause* voidaan perusmuodossaan ilmaista seuraavasti: Olkoon $X_1, X_2, \ldots, X_i, \ldots$ jono samoin jakautuneita riippumattomia satunnaismuuttujia, joiden (yhteistä) odotusarvoa merkitään $\mu = E(X_i)$ ja varianssia vastaavasti $\sigma^2 = \mathrm{Var}(X_i)$. Silloin jokaisella x:n reaaliarvolla

$$P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \le x\right) \to \Phi(x), \qquad \text{kun } n \to \infty.$$

(Vrt. kurssin *Johdatus todennäköisyyslaskentaan* luentomateriaali.) Huomaa, että vasen puoli tässä raja-arvotuloksessa voidaan kirjoittaa yhtä hyvin muotoon

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right),$$

jolloin raja-arvolauseen sisältämä tulos voidaan tulkita *normaaliapproksimaationa* n:stä riippumattomasta havainnosta lasketun aritmeettisen keskiarvon $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ jakaumalle (kertymäfunktiolle), kun tämä keskiarvomuuttuja edellä esitetyllä tavalla ensin "standardoidaan" vähentämällä siitä muuttujan odotusarvo μ ja sitten vielä jakamalla näin saatu erotus muuttujan keskihajonnalla $\sigma/\sqrt{n}$.

On huomattava, että keskeisen raja-arvolauseen muotoilu koskee nimenomaan kertymäfunktioita tai niiden avulla ilmaistavia todennäköisyyksiä, ei tiheysfunktioita. Siten esimerkiksi binomijakautuneet satunnaismuuttujat saavat aina vain kokonaislukuarvoja ja tätä vastaten myös kunkin muuttujan

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

jakauma on puhtaasti diskreetti, kaiken todennäköisyysmassan keskittyessä äärelliseen pistejoukkoon. Edellä esitetyn raja-arvotuloksen oikealla puolella esiintyvä standardinormaalijakauma taas on jatkuva ja sillä on jopa jatkuva tiheysfunktio.

Edellä esitetystä keskeisen raja-arvolauseen muotoilusta on olemassa lukuisia yleistyksiä, joissa esitettyjä oletuksia on väljennetty mahdollisesti sekä muuttujien X_i "samoinjakautuneisuuden" että niiden riippumattomuuden suhteen. Varsin tunnettu on suomalaisen J. Lindebergin 1920-luvulla esittämä muuttujien X_i variansseja koskeva ehto.

Esimerkki 2.4 *Eräässä kyselytutkimuksessa pyydettiin vastausta kysymykseen "Harrastatko lenkkeilyä?" Vastaajia oli yhteensä* 400*. Oletetaan, että Suomen aikuisväestöstä todellisuudessa 20 % harrastaa lenkkeilyä ja että kyselyyn poimittuja vastaajia voidaan riittävän tarkasti pitää tästä väestöstä poimittuna yksinkertaisena satunnaisotoksena.*

(a) Olkoon $\bar{X}_{400}$ otoksesta määritetty lenkkeilijöiden osuus. Mikä on silloin sen odotusarvo ja keskihajonta?

(b) Arvioi normaalijakauma-approksimaatiota käyttäen todennäköisyyttä sille, että $\bar{X}_{400}$ on välillä 0.18 *ja* 0.22*.*

(c) Kuinka suuri tulisi otoskoon olla, että edellä kohdassa (a) kysytty keskihajonta olisikin vain puolet siitä mitä edellä saatiin?

(Tenttitehtävä 9.5.2006)

Ratkaisu.

(a) Yksittäisen vastauksen X_i odotusarvo ja varianssi ovat $E(X_i) = 0.2$ ja $\mathrm{Var}(X_i) = 0.2 \times 0.8 = 0.16$. Havainnot oletettiin toisistaan riippumattomiksi, jolloin keskiarvon $\bar{X}_{400}$ odotusarvo ja varianssi voidaan laskea käyttäen kaavoja

$$E\left(\bar{X}_{400}\right) = E\left(\frac{1}{400}\sum_{i=1}^{400} X_i\right) = \frac{1}{400}\sum_{i=1}^{400} E(X_i) = \frac{400 \times 0.2}{400} = 0.2,$$

$$\mathrm{Var}\left(\bar{X}_{400}\right) = \frac{1}{400}\mathrm{Var}(X_1) = \frac{0.16}{400} = 0.0004.$$

Lenkkeilijöiden osuuden keskihajonnaksi saadaan varianssin perusteella $\sigma\left(\bar{X}_{400}\right) = \sqrt{\mathrm{Var}\left(\bar{X}_{400}\right)} = 0.02$.

(b) Keskeisen raja-arvolauseen mukaan lenkkeilijöiden osuus $\bar{X}_{400}$ noudattaa suunnilleen jakaumaa $N(0.2, 0.02^2)$. Näin ollen todennäköisyyttä $P\left(0.18 < \bar{X}_{400} < 0.22\right)$ voidaan arvioida normaalijakauman avulla:

$$\begin{aligned} P\left(0.18 < \bar{X}_{400} < 0.22\right) &= P\left(\frac{0.18 - 0.2}{0.02} < \frac{\bar{X}_{400} - 0.2}{0.02} < \frac{0.22 - 0.2}{0.02}\right) \\ &= P\left(\frac{\bar{X}_{400} - 0.2}{0.02} < 1\right) - P\left(\frac{\bar{X}_{400} - 0.2}{0.02} < -1\right) \\ &\approx \Phi(1) - \Phi(-1) \approx 0.841 - 0.159 = 0.682. \end{aligned}$$

(c) Mikäli vastauksia on yhteensä n kappaletta, lenkkeilijöiden osuuden $\bar{X}_n$ keskihajonta $\sigma(\bar{X}_n)$ voidaan laskea vastaavasti kuin (a)-kohdassa:

$$\sigma(\bar{X}_n) = \sqrt{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{0.16}{n}} = \frac{0.4}{\sqrt{n}} \leq 0.01.$$

Tämä epäyhtälö voidaan ratkaista n:n suhteen, jolloin saadaan $n \geq 1600$, ts. otoskoon tulee olla vähintään 1600, jotta keskihajonta olisi puolet siitä mitä (a)-kohdassa.

Samaan tulokseen voi päätyä myös suoremmin toteamalla, että koska keskiarvon keskihajonta on kääntäen verrannollinen otoskoon neliöjuureen, se pienenee puoleen aiemmasta arvostaan, kun otoskoko kasvaa nelinkertaiseksi.

2.5 Uskottavuusfunktioiden muodostaminen jatkuvien jakaumien tapauksessa

Aivan vastaavalla tavalla kuin aiemmin kaavassa (2.3), riippumattoman otannan tapauksessa havaintoja $X_1, X_2, \ldots, X_n$ vastaava uskottavuusfunktio voidaan muodostaa myös silloin kun muuttujat X_i ovat jatkuvia. Ainoa ero diskreetteihin jakaumiin verrattuna on, että nyt joudutaan pistetodennäköisyysfunktioiden $p_{X_i}(x_i;\theta)$ asemesta käyttämään vastaavia tiheysfunktioita, joita voidaan merkitä esimerkiksi $f_{X_i}(x_i;\theta)$. Otosta $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ vastaava uskottavuusfunktio voidaan siten lausua tulona

$$f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta). \tag{2.14}$$

Tässä kaavassa vasemmalla puolella oleva tiheysfunktio on siis määritelty satunnaismuuttujan $\boldsymbol{X}_n$ arvoalueessa (joka tavallisesti on n-ulotteinen reaaliavaruus tai jokin sen osa), kun taas oikealla puolella tekijöinä ovat yksittäisten havaintojen X_i tiheysfunktiot.

Esimerkki 2.5 *Olkoon X_1, X_2 riippumaton otos $Tas\,(0,\theta)$ jakaumasta, ts. jakaumasta, jonka tiheysfunktiolla on arvo $1/\theta$ välillä $[0,\theta]$ ja arvo 0 muualla (missä θ tulkitaan jakauman parametriksi). Määritä havaintotulosta (x_1, x_2) vastaava uskottavuusfunktion lauseke ja suurimman uskottavuuden estimaatti. (Tenttitehtävä 9.5.2006)*

Ratkaisu. Tutkitaan ensin pelkästään havaintoa X_1. Sen tiheysfunktio on

$$f_{X_1}(x;\theta) = \begin{cases} 1/\theta, & \text{kun } x \in [0,\theta] \\ 0 & \text{muulloin.} \end{cases}$$

Tuntematon parametrinarvo θ ei voi olla pienempi kuin havaittu arvo x_1, joten tätä vastaava uskottavuusfunktio on siis

$$L(\theta;x_1) = \begin{cases} f_{X_1}(x_1;\theta) = 1/\theta, & \text{kun } \theta \geq x_1 \\ 0, & \text{kun } \theta < x_1. \end{cases}$$

Uskottavuusfunktio saavuttaa suurimman arvonsa pisteessä x_1, joten suurimman uskottavuuden estimaatti on siten $\hat{\theta}(x_1) = x_1$.

Kahden havainnon X_1, X_2 yhteistiheysfunktio on

$$f_{X_1,X_2}(x_1,x_2;\theta) = f_{X_1}(x_1;\theta)\, f_{X_2}(x_2;\theta) = \begin{cases} 1/\theta^2, & \text{kun } x_1, x_2 \in [0,\theta] \\ 0 & \text{muulloin.} \end{cases}$$

Vastaavasti kuin yhden havainnon kohdalla ei tuntematon parametrinarvo θ voi olla pienempi kuin kumpikaan havaituista arvoista (x_1, x_2). Näitä vastaava uskottavuusfunktio on siis

$$L(\theta;x_1,x_2) = \begin{cases} f_{X_1,X_2}(x_1,x_2;\theta) = 1/\theta^2, & \text{kun } \theta \geq \max(x_1,x_2) \\ 0, & \text{kun } \theta < \max(x_1,x_2). \end{cases}$$

Tämä uskottavuusfunktio saa suurimman arvonsa pisteessä $\max(x_1,x_2)$, joten

$$\hat{\theta}(x_1,x_2) = \max(x_1,x_2).$$

Huomautus 2.6 *Tällä kurssilla ei tulla käsittelemään systemaattisesti moniulotteisten jatkuvien todennäköisyysjakaumien teoriaa tai tämän teorian kytkentöjä useamman muuttujan differentiaali- ja integraalilaskentaan. Tässä voidaan kuitenkin todeta, että ns. todennäköisyyslaskennan kertolaskusääntö pätee myös useampiulotteisten jakaumien tiheysfunktioille, joten voimme aina kirjoittaa kaavassa* (2.14) *vasemmalla puolella esiintyvän tiheysfunktion — silloinkin kun muuttujat X_i eivät ole riippumattomia — "ketjuttamalla" tulona*

$$f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i \mid X_1 = x_1, X_2 = x_2, ..., X_{i-1} = x_{i-1};\theta).$$

*Tulon tekijät ovat kaikki "*1*-ulotteisen satunnaismuuttujan" tiheysfunktioita, jotka vain on ilmaistu ehdollisina aiemmille havainnoille. Kaava* (2.14) *seuraa tästä, miksi kukin muuttuja X_i tulkitaan aina aiemmista havainnoista $X_1, X_2, \ldots, X_{i-1}$ riippumattomaksi, kunhan vain parametrin arvo θ on annettu.*

2.6 Normaalijakauman parametrien estimoinnista

Odotusarvon estimointi suurimman uskottavuuden menetelmällä riippumattoman otoksen tapauksessa, kun varianssi on tunnettu: Tarkastelemme seuraavassa aluksi yksinkertaisuuden vuoksi jossakin määrin keinotekoista tilannetta, jossa

jakauman varianssi σ^2 oletetaan tunnetuksi ja jossa ongelmaksi jää siten odotusarvon μ estimointi. Koska σ^2 oletettiin tunnetuksi, sille annetaan tässä juuri tämä tunnettu kiinteä arvo eikä sitä seuraavassa enää merkitä jakauman parametriksi. Näin ollen tarkastellaan otosta kuvaavaa tilastollista mallia $\boldsymbol{M} = \{f_{\boldsymbol{X}_n}(\cdot;\mu); -\infty < \mu < \infty\}$. Tässä otoksen (satunnaisvektorin) $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ todennäköisyystiheys saadaan oletetun riippumattomuusominaisuuden mukaisesti (vrt. edellä olevat kaavat (2.12) ja (2.14)) tulona

$$\begin{aligned} f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\mu) &= \prod_{i=1}^{n} f_{X_i}(x_i;\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &\propto \exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right\}, \end{aligned} \tag{2.15}$$

missä verrannollisuusmerkki "$\propto$" tarkoittaa nyt verrannollisuutta parametrin μ suhteen. Maksimikohdan löytämiseksi on helpointa tarkastella tämän uskottavuusfunktion logaritmia, jolloin tulokseksi saadaan (jälleen kerran!) aritmeettinen keskiarvo

$$\hat{\mu}(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n. \tag{2.16}$$

Edellä kohdassa (2.13) todettiin jo, että kun tämä estimaattori tulkitaan satunnaismuuttujaksi $\hat{\mu}(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$, sen jakauma on $N(\mu, \sigma^2/n)$. Se on siten sekä *harhaton* että *tarkentuva* μ:n estimaattori.

Varianssin estimointi suurimman uskottavuuden menetelmällä riippumattoman otoksen tapauksessa, kun odotusarvo on tunnettu: Tässä tapauksessa malliparametrina on vain normaalijakauman varianssi σ^2, joten tarkastelemme mallia $\boldsymbol{M} = \{f_{\boldsymbol{X}_n}(\cdot;\sigma^2); \sigma^2 > 0\}$. Uskottavuusfunktiolla on nyt lauseke

$$\begin{aligned} f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\sigma^2) &= \prod_{i=1}^{n} f_{X_i}(x_i;\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{2.17}$$

Derivoimalla tämän lausekkeen logaritmi parametrin σ^2 suhteen, asettamalla se nollaksi ja ratkaisemalla tästä σ^2, saadaan suurimman uskottavuuden estimaattorin lausekkeeksi

$$\hat{\sigma}^2(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{2.18}$$

Kyseessä on siis jälleen aritmeettinen keskiarvo, mutta nyt sovellettuna havaintojen x_i "neliöityihin poikkeamiin" niiden odotusarvosta μ. Jos jälleen tulkitsemme estimaattorin satunnaismuuttujaksi, voimme aiemman kaavan (2.10) perusteella kirjoittaa sen muotoon

$$\hat{\sigma}^2(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} Z_i^2,$$

missä muuttujat Z_i ovat riippumattomia ja $N(0,1)$-jakautuneita, ts. *standardinormaalimuuttujia*. Tässä syntyvän muuttujan $\sum_{i=1}^n Z_i^2$ jakaumaa kutsutaan *khii-toiseen jakaumaksi vapausastelukuna* n ja merkitään sitä symbolilla $\boldsymbol{\chi}^2(n)$ (vrt. gamma-jakaumaa koskeva huomautus sivulla 28). Koska $E(Z_i^2) = \mathrm{Var}(Z_i) = 1$, on selvää, että $E(\hat{\sigma}^2(\boldsymbol{X}_n)) = \sigma^2$, joten tämä σ^2:n estimaattori on *harhaton*. Sen varianssia tarkastelemalla voidaan helposti näyttää, että estimaattori on myös *tarkentuva*.

Odotusarvon ja varianssin estimointi suurimman uskottavuuden menetelmällä riippumattoman otoksen tapauksessa, kun sekä odotusarvo että varianssi ovat tuntemattomia: Tilanne, jossa sekä odotusarvo μ että varianssi σ^2 joudutaan estimoimaan otoksen avulla, on käytännön kannalta selvästi tavallisempi ja siten tärkeämpi kuin kaksi edellistä, joissa toinen parametreista oletettiin tunnetuksi. Vaikka nyt tarkastelemmekin mallia $\boldsymbol{M} = \{f_{\boldsymbol{X}_n}(\cdot;\mu,\sigma^2); -\infty < \mu < \infty, \sigma^2 > 0\}$, uskottavuusfunktiolla on sama lauseke kuin edellä (2.17):n oikealla puolella (ts. tapauksessa, jossa odotusarvo oli annettu ja estimoimme varianssia) kunhan vain verrannollisuus "$\propto$" nyt tulkitaan sekä μ:n että σ^2:n suhteen. Koska kuitenkin myös μ on estimoitava parametri, joudumme tehtävän ratkaisemiseksi derivoimaan uskottavuusfunktion (logaritmin) osittain kummankin parametrin suhteen. Asettamalla nämä osittaisderivaatat nolliksi, saamme kahden yhtälön ryhmän, jonka ratkaisu $(\hat{\mu}(\boldsymbol{x}_n), \hat{\sigma}^2(\boldsymbol{x}_n))$ on

$$\hat{\mu}(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n, \qquad \hat{\sigma}^2(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}(\boldsymbol{x}_n))^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2. \tag{2.19}$$

Satunnaismuuttujaksi tulkitun estimaattorin $\hat{\mu}(\boldsymbol{X}_n)$ jakaumaa ja ominaisuuksia selvitettiin jo edellä. Sen sijaan kysymys estimaattorin $\hat{\sigma}^2(\boldsymbol{X}_n)$ jakaumasta on mutkikkaampi. Tällä kurssilla esitämme vain seuraavan tärkeän tuloksen ilman todistusta:

Lause 2.7 *Estimaattorit $\hat{\mu}(\boldsymbol{X}_n) = \bar{X}_n$ ja $\hat{\sigma}^2(\boldsymbol{X}_n)$ ovat riippumattomia todennäköisyyden $P_{(\mu,\sigma^2)}$ suhteen. Lisäksi*

$$\begin{aligned} \hat{\mu}(\boldsymbol{X}_n) &\quad \textit{jakautuu kuten } N(\mu, \sigma^2/n)\textit{-muuttuja ja} \\ \hat{\sigma}^2(\boldsymbol{X}_n) &\quad \textit{kuten } \frac{\sigma^2}{n}\boldsymbol{\chi}^2(n-1)\textit{-muuttuja.} \end{aligned}$$

Tässä tuloksessa on kaksi yllättävää seikkaa: Ensiksi, estimaattorit ovat riippumattomia siitäkin huolimatta, että edellinen (havaintojen aritmeettinen keskiarvo) esiintyy jälkimmäisen lausekkeessa. Toiseksi, jälkimmäisen $\boldsymbol{\chi}^2$-muuttujan vapausasteluku on yhtä pienempi kuin edellä tapauksessa (2.18). Näin ollen suurimman uskottavuuden estimaattori $\hat{\sigma}^2(\boldsymbol{X}_n)$ ei myöskään ole harhaton: $E_{(\mu,\sigma^2)}(\hat{\sigma}^2(\boldsymbol{X}_n)) = \frac{n-1}{n}\sigma^2$. Otoskoon n kasvaessa harhan määrä (odotusarvon mielessä) luonnollisesti pienenee, koska $\frac{n-1}{n} \to 1$. Varsin usein kuitenkin käytetään estimaattorin $\hat{\sigma}^2(\boldsymbol{x}_n)$ (kaava (2.19)) asemesta kaavan

$$s^2 = s^2(\boldsymbol{x}_n) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 \tag{2.20}$$

määrittelemää varianssiestimaattoria, joka satunnaismuuttujaksi tulkittuna on harhaton.

Huomautus 2.8 *Huomaa kuitenkin, ettei s^2:n neliöjuuri $s = s(\boldsymbol{X}_n)$ ole enää harhaton σ:n estimaattori! Tämä johtuu siitä, että neliöjuuren otto ei laskutoimituksena ole lineaarinen, joten sen ja odotusarvon järjestystä ei voi vaihtaa.*

Pohdintaa. Edellä olevat esimerkit kuvaavat sitä verraten suoraviivaista menettelyä, jolla suurimman uskottavuuden estimaattorin lauseke voidaan johtaa yksinkertaisten parametristen tilastollisten mallien tapauksessa. Tarkastelluissa tapauksissa oli myös mahdollista löytää täsmällinen muoto estimaattorin ns. *otosjakaumalle*, ts. sen jakaumalle kun otos tulkitaan satunnaismuuttujaksi annetun tilastollisen mallin puitteissa. Monimutkaisemmissa ja sen vuoksi analyyttisesti vaikeammin hallittavissa tilanteissa mallin mukaisen uskottavuusfunktion maksimointitehtävä joudutaan kuitenkin usein suorittamaan numeerisesti, koska sille ei ole suljetussa muodossa olevaa ratkaisua. Toisaalta tilastotieteen teoriassa on runsaasti matemaattisia tuloksia, jotka kuvaavat estimaattoreiden asymptoottisia ominaisuuksia otoskoon n kasvaessa. Suurimman uskottavuuden estimaattorit ovat — varsin yleisten oletusten vallitessa — likimäärin normaalisesti jakautuneita ja myös likimäärin harhattomia, kun otoskoko on suuri. Tästä johtuen suurimman uskottavuuden estimaateille käytetään usein näihin estimaattoreiden asymptoottisiin ominaisuuksiin perustuvia *luottamusvälejä*. Luottamusvälien käsitteeseen palaamme vielä myöhemmin.

Harjoitustehtävä 2.6 *Erään sähkölampputyypin paloajan on todettu noudattavan likimäärin normaalijakaumaa, jonka odotusarvo on* 1000 *h ja keskihajonta* 200 *h. Valmistuserästä tehdään yksinkertainen* 100 *kappaleen satunnaisotos ja siihen kuuluvien lamppujen palamisajat mitataan. Kun niiden keskiarvoa merkitään* $\bar{X}_{100}$*, määritä todennäköisyys* $P(980 < \bar{X}_{100} < 1020)$*. Etsi myös sellainen luku* b*, jolle pätee* $P(1000 - b < \bar{X}_{100} < 1000 + b) = 0.99$*. (Tenttitehtävä 7.3.2006)*

Harjoitustehtävä 2.7 *Eräässä kokeessa pyrittiin selvittämään kahden koostumukseltaan erilaisen rehun vaikutusta broilerien painoon. Tätä varten satunnaistettiin 40 kasvatusbroileria kahteen 20 broilerin ryhmään, josta 1. ryhmää ruokittiin rehulla A ja 2. ryhmää rehulla B. Ruokintakokeen päättymisen jälkeen kummastakin ryhmästä mitattiin broilerien kokeen aikana kertynyt painonlisäys, merkiten havaittuja ryhmäkeskiarvoja* $\bar{X}_{1,20}$ *ja* $\bar{X}_{2,20}$*. Käytetystä satunnaistamisesta johtuen arvioitiin, että näitä muuttujia voidaan perustellusti pitää riippumattomina.*

(a) Oletetaan sitten, että rehulla A ruokittujen broilereiden painonlisäys olisi jakautunut likimäärin kuten $N(360g, (55g)^2)$ *ja vastaavasti rehulla B ruokittujen painonlisäys kuten* $N(385g, (50g)^2)$*. Mitkä ovat muuttujien* $\bar{X}_{1,20}$ *ja* $\bar{X}_{2,20}$ *jakaumat? Entä erotuksen* $\bar{X}_{1,20} - \bar{X}_{2,20}$ *jakauma?*

(b) Mikä on todennäköisyys sille, että rehulla B ruokittujen broilereiden painolisäys on vähintään 25 g suurempi kuin rehulla A ruokittujen?

Harjoitustehtävä 2.8 *Johda edellä kaavassa* (2.16) *annettu normaalijakauman odotusarvon suurimman uskottavuuden estimaattorin lauseke.*

Harjoitustehtävä 2.9 *Eräässä vaarallisessa risteyksessä tapahtuvien liikenneonnettomuuksien lukumäärän havaittiin vaihtelevan viikoittain siten, että lukumäärän odotusarvo oli* 2.2 *ja keskihajonta* 1.4*. (Huomaa, ettei viikon aikana tapahtuvien onnettomuuksien lukumäärä voi satunnaismuuttujaksi tulkittuna tietenkään olla normaalijakautunut — jo siitäkin syystä, että se saa vain kokonaislukuarvoja* $0, 1, 2, \ldots$*.) Olkoon sitten erään vuoden* 52 *viikon aikana havaittujen viikoittaisten lukumäärien keskiarvo* $\bar{X}_{52}$*.*

(a) Minkä approksimaation keskeinen raja-arvolause antaa muuttujan $\bar{X}_{52}$ *jakaumalle?*

(b) Mikä on tällä perusteella saatava likiarvo todennäköisyydelle $P(\bar{X}_{52} < 2)$*?*

(c) Määritä samalla tavalla likiarvo todennäköisyydelle

$$P(\textit{onnettomuuksien lukumäärä vuoden aikana} < 100).$$

2.7 Parametriestimointi Bayes-viitekehyksessä

Tämän kappaleen tarkoituksena on, paitsi demonstroida Bayes-paradigmaan perustuvaa estimointimenetelmää parin esimerkin kautta, kiinnittää huomiota niihin käsitteellisiin eroihin, joita edellä esitettyjen uskottavuusfunktioon perustuvien mutta luonteeltaan frekventististen ja toisaalta subjektiiviseen todennäköisyyteen perustuvien Bayes-menetelmien välillä on. Nämä lähestymistavat ovat tietyssä suhteessa ikään kuin toistensa peilikuvia. Suurimman uskottavuuden estimaattorien tapauksessa satunnaisuus, samoin kuin tähän liitettävät todennäköisyystarkastelut, liitetään havaintojen $\boldsymbol{X}_n$ *potentiaaliseen käyttäytymiseen* parametrin arvon θ pysyessä kiinteänä. Bayes-päättely sen sijaan on luonteeltaan ehdollista siinä mielessä, että siinä alun pitäen satunnaismuuttujaksi tulkittu $\boldsymbol{X}_n$ kiinnitetään sen havaittuun arvoon $\boldsymbol{x}_n$. Sen sijaan tuntematon parametri θ tulkitaan satunnaismuuttujaksi, jonka satunnaisuus vastaa kirjaimellisesti sitä, ettei sen arvoa tunneta. (Huom. Parametrien kohdalla tulkinnallista eroa ei ole tapana korostaa käyttämällä eri merkintöjä, esim. θ ja Θ, vaan molemmissa tapauksissa parametria merkitään yksinkertaisesti θ:lla.) Vastaavasti todennäköisyys ymmärretään kvantitatiiviseksi ilmaukseksi henkilön epävarmuudesta koskien parametrin todellista arvoa; täten todennäköisyyksille annetut numeeriset arvot voivat vaihdella henkilöstä toiseen ja riippuvat myös siitä, mitä informaatiota näillä on käytettävissään.

Jos lähtökohdaksi valitaan todennäköisyyden subjektiiviseen tulkintaan perustuva Bayes-paradigma, voidaan tilastollisen mallin $\boldsymbol{M}^* = \{p(\theta, \boldsymbol{x}_n)\}$ puitteissa Bayesin kaavan

$$p(\theta|\boldsymbol{x}_n) \propto p(\theta)\, f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\theta) \tag{2.21}$$

ja datan perusteella määritettyä *posteriorijakaumaa* pitää tietyssä mielessä tilastollisen estimointiongelman kaiken kattavana ratkaisuna: Se kertoo suoraan, miten tuntemattomaan parametriin θ liitettäviä (*priori*)*todennäköisyyksiä* päivitetään, kun saadaan uutta tietoa havainnon $\boldsymbol{X}_n = \boldsymbol{x}_n$ muodossa. Tässä tiheysfunktio $f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\theta)$ vastaa edellä mallien $\boldsymbol{M}$ puitteissa käsiteltyä uskottavuusfunktiota $f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta)$; erona on vain se, että koska θ tulkitaan mallissa $\boldsymbol{M}^*$ satunnaismuuttujaksi, myös havainnon $\boldsymbol{X}_n$ jakauman riippuvuus θ:n arvosta ymmärretään ehdolliseksi todennäköisyydeksi eikä vain riippuvuudeksi parametrin arvosta. Huomaa myös, että kaavassa (2.21) esiintyvä priorijakauma $p(\theta)$ on mallin $\boldsymbol{M}^*$ osa, sillä se määrittyy *reunajakaumana* muuttujien θ ja $\boldsymbol{X}_n$ yhteisjakaumasta $p(\theta, \boldsymbol{x}_n)$.

Erityisesti, jos olemme kiinnostuneet muotoa "Kuuluuko parametrin oikea arvo johonkin tiettyyn parametriavaruuden osajoukkoon A?" olevasta kysymyksestä, voimme aina päätyä vastaavaan ehdolliseen todennäköisyyteen $P(\theta \in A|\boldsymbol{X}_n = \boldsymbol{x}_n)$ yksinkertaisesti integroimalla posterioritiheysfunktiota tarkasteltavan parametriavaruuden osajoukon A yli, ts.

$$P(\theta \in A \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = \int_A p(\theta|\boldsymbol{x}_n)\, d\theta. \tag{2.22}$$

Diskreettien jakaumien tapauksessa vastaava tulos saadaan laskemalla yhteen (posteriori)pistetodennäköisyydet niiden θ:n arvojen osalta, jotka kuuluvat joukkoon A.

Jos parametri on yksiulotteinen reaaliarvoinen muuttuja, on usein on kiinnostavaa tarkastella kysymyksiä, jotka ovat esimerkiksi muotoa "Onko θ välillä (θ_1, θ_2)?" tai "Onko $\theta > \theta_1$?" Näihin kysymyksiin ei tietenkään voi saada varmaa ja aina oikeaa "kyllä"- tai "ei"-vastausta, koska tilastolliseen päättelyyn — sen luonteen mukaisesti — sisältyy aina epävarmuutta. Sen sijaan kaavaa (2.22) soveltamalla voidaan laskea (posteriori)-todennäköisyys sille, että kysymyksen sisältämä vaihtoehto pitää paikkansa. Edellisessä tapauksessa on vain valittava tarkasteltavaksi joukoksi väli $A = (\theta_1, \theta_2)$, jälkimmäisessä taas puoliakseli $A = (\theta_1, \infty)$. Jos parametri θ on useampiulotteinen (vektoriarvoinen) ja tarkasteltava kysymys kuitenkin kohdistuu johonkin tiettyyn sen koordinaattiin, voidaan tuloksena saadusta posterioritiheydestä aina periaatteessa integroida muut dimensiot pois, ts. voidaan tarkastella kiinnostuksen kohteena olevan suureen reunajakaumaa.

Yleisesti ottaen posteriorijakauman avulla voidaan laskea muotoa

$$E(g(\theta) \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = \int g(\theta)\, p(\theta|\boldsymbol{x}_n)\, d\theta \tag{2.23}$$

olevia *posterioriodotusarvoja* jollekin mielenkiinnon kohteena olevalle parametrin funktiolle g. Valitsemalla erityisesti funktio g identtiseksi kuvaukseksi $g(\theta) = \theta$, saadaan tästä määritetyksi parametrin θ oma odotusarvo $E(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$. Jos taas valitaan $g(\theta) = \theta^2$, päädytään toiseen momenttiin $E(\theta^2 \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$ ja tätä kautta edelleen posteriorivariansiin

$$\mathrm{Var}(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = E(\theta^2 \mid \boldsymbol{X}_n = \boldsymbol{x}_n) - [E(\theta \mid \boldsymbol{X}_n = \boldsymbol{x}_n)]^2.$$

Huomaa myös, että odotusarvon kaava (2.23)) palautuu todennäköisyyden kaavaksi (2.22), jos g valitaan joukon A indikaattorifunktioksi (ts. määritellään se siten, että $g(\theta) = 1$ kun $\theta \in A$ ja $g(\theta) = 0$ kun $\theta \notin A$).

Posteriorijakauma määräytyy näin annetun mallin puitteissa aina yksikäsitteisesti havaintojen $\boldsymbol{X}_n = \boldsymbol{x}_n$ perusteella. Toisaalta malli $\boldsymbol{M}^* = \{p(\theta, \boldsymbol{x}_n)\}$ määritellään käytännössä valitsemalla tulon $p(\theta, \boldsymbol{x}_n) = p(\theta) f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\theta)$ kumpikin tekijä erikseen, ts. *priorijakauma* $p(\theta)$ ja *uskottavuusfunktio* $f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\theta)$. Bayesin kaavan soveltaminen tästä lähtökohdasta, aina todennäköisyyksien numeerista määrittämistä myöten, voi kuitenkin olla käytännöllisistä syistä hankalaa. Näin on laita varsinkin tilanteissa, joissa parametriavaruuden dimensio on hyvin suuri. Toisaalta, kuten aiemmin todettiin Binomi- ja Beta-jakaumien yhteydessä (kappale 1.6), uskottavuusfunktion ja sitä vastaavan priorin valinnat voidaan tietyissä tilanteissa sovittaa yhteen siten, että posteriorin määritystehtävä on hallittavissa helposti myös analyyttisin keinoin.

Esimerkki 2.9 *Olkoon $X_1, X_2, \ldots, X_n$ riippumaton otos tiheysfunktion $f_X(x;\theta)$ mukaan jakautuneita satunnaismuuttujia ja olkoon $\hat{\theta} = \hat{\theta}(\boldsymbol{x}_n)$ niiden havaittuja arvoja $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ vastaava suurimman uskottavuuden estimaatti. Tarkastellaan sitten vastaavaa parametriestimointitehtävää Bayes-päättelyn näkökulmasta olettaen, että θ:lle on määritelty prioriksi tasajakauma välillä (a, b). Näytä, että jos $\hat{\theta}$ osuu välille (a, b), se on myös "todennäköisin θ:n arvo" (posterioritodennäköisyyden mielessä), ts. posterioritiheysfunktio saavuttaa suurimman arvonsa juuri pisteessä $\hat{\theta}$. (Tenttitehtävä 14.11.2007)*

Ratkaisu. Määritelmänsä mukaan $\hat{\theta}$ on se arvo, joka maksimoi havaintoja $\boldsymbol{x}_n = (x_1, x_2, ..., x_n)$ vastaavan uskottavuusfunktion

$$L(\theta; \boldsymbol{x}_n) = f_{\boldsymbol{X}_n}(\boldsymbol{x}_n; \theta)$$

parametrin θ suhteen. Havaintoja $\boldsymbol{x}_n$ vastaavaksi posterioritiheysfunktioksi saadaan kaavan (2.21) mukaisesti

$$p(\theta|\boldsymbol{x}_n) \propto p(\theta) f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta) = \begin{cases} \frac{1}{b-a} f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta), & \text{kun } \theta \in (a,b) \\ 0 & \text{muulloin.} \end{cases}$$

Koska tekijä $\frac{1}{b-a}$ ei riipu θ:sta, niin posteriorijakauman mielessä todennäköisin θ:n arvo löydetään maksimoimalla $f_{\boldsymbol{X}_n}(\boldsymbol{x}_n;\theta)$ välillä (a,b). Mikäli suurimman uskottavuuden estimaatti $\hat{\theta}$ on tällä välillä, niin posterioritiheysfunktiokin saavuttaa suurimman arvonsa pisteessä $\hat{\theta}$.

Odotusarvon estimointi Bayes-menetelmällä riippumattoman normaalijakautuneen otoksen tapauksessa, kun varianssi on tunnettu: Estimointiongelmana tämä vastaa täysin aikaisemmin käsiteltyä vastaavaa tehtävää, joten sen lähtökohtana voidaan pitää kaavassa (2.15) annettua uskottavuusfunktion lauseketta

$$f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu) \propto \exp\left\{ \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2} \right\}, \tag{2.24}$$

Ainoana erona aiempaan on se, että nyt tulkitsemme oikean puolen ehdolliseksi todennäköisyystiheydeksi ja siksi käytämme merkintää $f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu)$. Sen jälkeen parametrin μ havaintoja $\boldsymbol{x}_n$ vastaava posteriorijakauma määräytyy Bayesin kaavan

$$p(\mu|\boldsymbol{x}_n) \propto p(\mu) f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu) \tag{2.25}$$

mukaisesti. Jotta kaavaa voitaisiin soveltaa, täytyy vielä valita μ:lle jokin sopiva priorijakauma $p(\mu)$. Osoittautuu, että jos myös tämä valinta tehdään normaalijakaumaperheestä, on valitulla priorilla ja uskottavuusfunktiolla analyyttisen ratkaisun löytämistä helpottava *liittojakaumaominaisuus*. Tämä nähdään seuraavasti: Olkoon μ:n priorijakauman tiheysfunktio

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{ -\frac{(\mu-\mu_0)^2}{2\sigma_0^2} \right\} \propto \exp\left\{ -\frac{\mu^2}{2\sigma_0^2} + \frac{\mu_0\mu}{\sigma_0^2} \right\}, \tag{2.26}$$

missä μ_0 ja σ_0^2 ovat valitun μ:n priorijakauman odotusarvo ja varianssi. Bayes-päättelyn periaatteiden mukaisesti tämä valinta kuvastaa henkilön (esimerkiksi "minun") käsitystä oikeasta μ:n arvosta. Kuten edellä todettiin, tilastotieteessä on varsin tavallista viitata tällöin ajateltavissa olevaan äärettömään perusjoukkoon tai populaatioon ja tulkita μ tämän populaation ominaisuudeksi. Suuri varianssin σ_0^2 arvo vastaa tilannetta, jossa μ:tä koskeva ennakkotieto on heikkoa, pieni varianssi taas päinvastaista tilannetta. Huomaa siis, että vaikka parametreille μ ja σ^2 voidaan myös Bayes-päättelyssä antaa otantaan liittyvä frekventistinen tulkinta, näin ei voi enää tehdä priorijkauman parametrien μ_0 ja σ_0^2 kohdalla (ainakaan ilman ajatusta "populaatioiden populaatiosta" ja sieltä tapahtuvasta kaksivaiheisesta otannasta).

Yhdistämällä nyt kaavojen (2.24) ja (2.26) sisältämä tieto Bayesin kaavan (2.25) osoit-

tamalla tavalla saamme μ:n posterioritiheyden lausekkeeksi (vakiotekijää vaille)

$$\begin{aligned} p(\mu|\boldsymbol{x}_n) &\propto \exp\left\{-\frac{\mu^2}{2\sigma_0^2}+\frac{\mu_0\mu}{\sigma_0^2}\right\}\exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i-\frac{n\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^n x_i\right)\mu\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left[\mu-\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^n x_i\right)\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right]^2\right]\right\} \end{aligned}$$

Oikealla puolella oleva lauseke on kuitenkin (jälleen vakiotekijää vaille) sellaisen normaalijakauman tiheysfunktio, jonka odotusarvo on $\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^n x_i\right)\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$ ja varianssi $\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$.

Johdettu tulos tulee havainnollisempaan muotoon, jos merkitään $\omega = 1/\sigma^2$ ja $\omega_0 = 1/\sigma_0^2$ ja vielä $\sum_{i=1}^n x_i = n\bar{x}_n$ Parametreja ω ja ω_0, jotka siis ovat varianssien käänteislukuja, on tapana kutsua jakauman $f_{\boldsymbol{X}_n}$ ja priorijakauman p *tarkkuudeksi* (engl. *precision*). Näitä merkintöjä käyttäen voimme todeta posterioritiheyden olevan muotoa

$$p(\mu|\boldsymbol{x}_n) \propto \exp\left\{-\frac{\omega_0+n\omega}{2}\left(\mu-\frac{\omega_0\mu_0+n\omega\bar{x}_n}{\omega_0+n\omega}\right)^2\right\}.$$

Se on normaalijakauma, jonka odotusarvo on $\frac{\omega_0\mu_0+n\omega\bar{x}_n}{\omega_0+n\omega}$ ja tarkkuus vastaavasti $\omega_0+n\omega$. Posteriorijakauman odotusarvo on näin ollen painotettu keskiarvo priorijakauman odotusarvosta μ_0 ja otoksen sisältämien havaintojen aritmeettisesta keskiarvosta $\bar{x}_n$, kun niitä vastaavina painokertoimina ovat priorijakauman tarkkuus ω_0 ja havaintojen jakauman tarkkuus ω painotettuna edelleen otoksen koolla eli havaintojen lukumäärällä.

Pohdintaa. Tarkastelemme nyt lähemmin edellä johdettua μ:n posteriorijkaumaa $p(\mu|\boldsymbol{x}_n)$. Havaittiin, että posteriorijakauman tarkkuusparametrin arvo saadaan laskemalla yhteen priorijakauman tarkkuus ω_0 ja otoksesta lasketun aritmeettisen keskiarvon $\bar{X}_n$ jakauman tarkkuus $n\omega$. (Tässä voi palauttaa mieleen aiemmin johdetun tuloksen (2.13), jonka mukaan n:n havainnon aritmeettisen keskiarvon varianssi on $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$ ja siis vastaava tarkkuusparametrin arvo $(\sigma^2/n)^{-1} = n/\sigma^2 = n\omega$.) Odotusarvoa μ estimoitaessa "yhdistetyn tarkkuusparametrin" arvo kasvaa siten lineaarisesti otoskoon kasvaessa. Vastaavasti nähdään, että posterioriodotusarvon kaavassa $E(\mu|\boldsymbol{x}_n) = \frac{\omega_0\mu_0+n\omega\bar{x}_n}{\omega_0+n\omega}$ otoskeskiarvon $\bar{x}_n$ painokerroin kasvaa otoskoon mukana, kun taas prioriodotusarvon μ_0 painokerroin pysyy muuttumattomana. Näin siis riittävän suurilla otoksilla otoskeskiarvo voittaa prioriodotusarvon posterioriodotusarvon määrittäjänä prioriodotusarvon.

Tätä ajatusta voidaan koettaa vielä täsmentää seuraavasti: Oletetaan, että päättymätön jono $X_1, X_2, \ldots, X_n, \ldots$ satunnaismuuttujia olisi poimittu riippumattomasti normaalijakaumasta, jonka odotusarvo on μ_1 ja varianssi jo edellä mainittu tunnettu σ^2. (Tällainen jono voitaisiin periaatteessa konstruoida tietokonesimulaationa, jossa μ_1 ja σ^2 toimisivat simulaatioparametreina. Ajatus vastaa luentokurssin alussa esitettyä ajatusta todennäköisyysmallista dataa generoivana mekanismina.) Suurten lukujen lain perusteella tiedetään silloin, että vastaavat otoskeskiarvot $\bar{X}_n$ konvergoivat kohti odotusarvoa μ_1. (Huomaa, että koska otoskeskiarvot $\bar{X}_n$ ovat satunnaismuuttujia, suurten lu-

kujen lain täsmällisempi matemaattinen muotoilu edellyttäisi viittausta johonkin todennäköisyyteen, jonka suhteen konvergenssi sitten tapahtuu. Viitetodennäköisyytenä on luonnollisesti se jakauma, josta jono $X_1, X_2, \ldots, X_n, \ldots$ oletettiin poimituksi, tässä siis $N(\mu_1, \sigma^2)$.) Sijoittamalla tämä tulos edellä johdettuun μ:n posterioriodotusarvon $E(\mu|\boldsymbol{X}_n)$ lausekkeeseen nähdään heti, että myös tämä jono konvergoi otoskoon n kasvaessa kohti "oikeata odotusarvoa" μ_1. Koska samalla myös posteriorijakauman $p(\mu|\boldsymbol{X}_n)$ tarkkuus kasvaa mielivaltaisen suureksi (posteriorivarianssin lähestyessä vastaavasti nollaa), voidaan todeta, että myös nämä jakaumat konvergoivat (sopivasti määritellyn jakaumakonvergenssin mielessä) kohti pisteeseen μ_1 keskittynyttä ykkösen suuruista pistetodennäköisyyttä. Voidaan siis sanoa, että otoskoon kasvaessa rajatta oikea parametrin arvo μ_1 löytyy lopulta riippumatta siitä, miten priorijakauman parametrit μ_0 ja σ_0^2 on valittu. Tätä tulosta voidaan pitää eräässä mielessä aiemman, odotusarvon suurimman uskottavuuden estimaattorin $\hat{\mu}(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$ tarkentuvuutta koskevan tuloksen Bayes-paradigman mukaisena vastineena. Sen täsmällinen matemaattinen muotoilu on hankalampi lähinnä siitä syystä, että se koskee kokonaisten (posteriori)jakaumien rajakäyttäytymistä, ei vain piste-estimaattien $\hat{\mu}(\boldsymbol{X}_n)$.

Tarkastellaan sitten hieman toisenlaista tilannetta, missä otos ja myös prioriodotusarvo μ_0 pysyvät kiinteinä, mutta prioritarkkuus ω_0 lähestyy nollaa. Tämä vastaa tilannetta, jossa todennäköisyysarviot tekevä henkilö ("minä") haluaa ikään kuin etäännyttää itsensä μ:n estimointia koskevasta tehtävästä ja "antaa vain datan puhua". Tällöin nähdään heti edellä posterioriodotusarvolle ja posteriorivarianssille johdetuista lausekkeista $E(\mu|\boldsymbol{x}_n) = (\omega_0\mu_0 + n\omega\bar{x}_n)(\omega_0 + n\omega)^{-1}$ ja $\mathrm{Var}(\mu|\boldsymbol{x}_n) = (\omega_0 + n\omega)^{-1}$, että vastaavana raja-arvona saatava posteriorijakauma $p(\mu|\boldsymbol{x}_n)$ on normaalijakauma $N(\bar{x}_n, \sigma^2/n)$, jonka odotusarvo on $\bar{x}_n$ ja varianssi σ^2/n. Tätä tulosta tarkasteltaessa on tärkeää huomata, että nyt on kysymys parametrin μ jakaumasta kun havainnot $\boldsymbol{X}_n = \boldsymbol{x}_n$ on annettu, ei siis esimerkiksi otoskeskiarvon $\bar{X}_n$ jakaumasta $N(\mu, \sigma^2/n)$, kun jakauman odotusarvo μ on annettu. Voidaan sanoa, että näissä kahdessa jakaumassa μ:n ja $\bar{X}_n$:n roolit ovat vaihtuneet päinvastaisiksi. Molemmissa tapauksissa kuitenkin jakaumien varianssi on sama σ^2/n, joten ne ovat normaalijakaumina saman muotoisia.

Tämä tarkastelu johtaa meidät kysymään, voitaisiinko tällainen eräänlaisena posteriorijakaumien raja-arvona saatava tulos johtaa myös suoraan, ts. onko mahdollista valita sellainen μ:n priorijakauma, joka yhdistettynä edellä tarkasteltuun normaalijakauman uskottavuusfunktioon (2.24) tuottaisi μ:lle posteriorijakauman $N(\bar{x}_n, \sigma^2/n)$. Vastaus on osittain myönteinen, osittain kielteinen. Avaimen siihen saa tarkastelemalla mitä priorijakauman $N(\mu_0, \sigma_0^2)$ tiheysfunktiolle tapahtuu, kun $\sigma_0^2 \to \infty$: Nähdään helposti, että tiheysfunktion arvot lähestyvät nollaa kaikissa reaaliakselin pisteissä. Toisaalta, verrattaessa tiheysfunktioiden arvoja missä tahansa kahdessa pisteessä, niiden suhde lähestyy lukua 1, joten prioritiheyden raja-arvoksi tulisi vakio koko reaaliakselilla. Edellinen ajatus johtaa ehdotukseen käyttää prioritiheysfunktiota, joka olisi identtisesti nolla. Tähän kuitenkin sisältyy se ilmeinen ongelma, ettei kyse silloin ole enää mistään todennäköisyysjakaumasta; myös Bayesin kaava supistuisi triviaaliin ja hyödyttömään muotoon "$0 \propto 0$". Jotta tämä tilanne voitaisiin välttää, voidaan sitten ajatella, että kokeiltaisiin μ:n prioritiheydelle jotakin positiivista vakioarvoa $p(\mu) \equiv c > 0$. Tällaista prioritiheysfunktiota voitaisiin pitää siinä mielessä epäinformatiivisena, että se ei aseta mitään μ:n arvoa erityisasemaan muihin verrattuna. Se ei kuitenkaan voi olla minkään todennäköisyysjakauman tiheysfunktio, sillä sen integraali yli koko reaaliakselin olisi ääretön.

Varsinkin aiemmin oli tavallista soveltaa Bayesin kaavaa myös sellaisissa tapauksissa,

joissa itse priorijakauma ei ollut integroituva, mutta joissa Bayesin kaavassa nimittäjässä oleva verrannollisuustekijä kuitenkin on äärellinen. Esillä olevassa tapauksessa voidaan helposti todeta, että $\int_{-\infty}^{\infty} f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu)\, d\mu < \infty$, kun uskottavuusfunktio valitaan kaavan (2.24) mukaisella tavalla. Integraali luonnollisesti säilyy äärellisenä, jos integroitavana oleva uskottavuusfunktio kerrotaan positiivisella vakioarvolla c. Tuloksena saatavaa Bayesin kaavaa

$$p(\mu|\boldsymbol{x}_n) \propto p(\mu) f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu) \propto f_{\boldsymbol{X}_n}(\boldsymbol{x}_n|\mu)$$

voidaan myös tässä tilanteessa pitää järkevänä ainakin siinä mielessä, että se määrittelee μ:n todennäköisyysjakauman. Toisaalta on esitetty varsin painavia perusteita myös tällaista epä-integroituvien priorijakaumien käyttöä vastaan. Voidaan myös kysyä, onko vakiofunktio koko reaaliakselilla todella järkevä vastine joskus tavoitellulle epäinformatiivisuudelle: koko reaaliakselin yli lasketusta integraalista, joka on ääretön, tulee minkä tahansa äärellisen välin osalle vain äärellinen — ja siis suhteessa merkityksetön — osuus.

Luku 3

Luottamusvälit ja luottamusjoukot

Kuten aiemmassa luentotekstissä todettiin, suurimman uskottavuuden menetelmällä saatavien piste-estimaattien avulla pyritään löytämään parametriavaruudesta yksi hyvä arvaus oikealle parametrin arvolle, ts. piste, jonka voitaisiin ajatella olevan (useimmiten tosin vain kuvitteellisen) dataa generoivan mekanismin oikea parametrin arvo. Pistemäinen arvaus kuitenkin sattuu vain hyvin harvoin juuri oikeaan kohtaan. Sen vuoksi *väliestimoinnilla* pyritään löytämään tämän pistemäisen arvion ympärille eräänlainen toleranssialue, johon sitten voitaisiin perustellummin ajatella oikean parametrin arvon kuuluvan. Frekventistisen lähestymistavan puitteissa ratkaisuna tähän tehtävään esitetään tavallisesti ns. *luottamusvälin* (engl. *confidence interval*) käsite.

Pohdintaa. Jos parametriavaruus on useampiulotteinen, voidaan siinäkin tapauksessa pyrkiä löytämään saadun piste-estimaatin ympäriltä jokin geometrialtaan sopivasti määritelty parametriavaruuden osajoukko, jolla olisi vastaava toleranssialueominaisuus. Toisaalta vektoriarvoisten parametrien tapauksessa on varsin tavallista, että mielenkiinto kohdistuu erityisesti niiden yhteen tai muutamaan koordinaattiin. Silloin on luonnollista pyrkiä eliminoimaan mallin muiden ns. *kiusaparametrien* (engl. *nuisance parameter*) vaikutus varsinaisesti kiinnostuksen kohteena olevia parametrin koordinaatteja koskeviin johtopäätöksiin. Tässä tarkoituksessa kiusaparametrit pyritään usein eliminoimaan korvaamalla ne jollakin tavalla, esim turvautumalla ns. *profiiliuskottauuden* käsitteeseen.

Suoraviivainen — tosin toistaiseksi suhteellisen harvoin käytetty — tapa väliestimointiongelman ratkaisemiseksi perustuu suoraan uskottavuusfunktion arvojen vertailuun parametriavaruuden eri osissa. Suurimman uskottavuuden estimaatti on uskottavuusfunktion huippukohta ja sen välittömässä ympäristössä olevat arvot myös vastaavassa mielessä uskottavampia saadun havainnon selittäjinä kuin muut parametriarvot. Jos uskottavuusfunktio on yksihuippuinen ja parametriavaruus reaaliakseli tai jokin sen väli, suurimman uskottavuuden estimaatin ympärille voidaan aina erottaa väli, jonka sisällä olevat arvot ovat uskottavampia kuin ulkopuolella olevat. Useampiulotteisen parametriavaruuden tapauksessa tilannetta voidaan tavallisesti kuvata topografikartan tapaan uskottavuusfunktion tasa-arvokäyrillä, joiden sisään jäävillä alueilla on vastaava ominaisuus. Vaikka tällainen uskottavuusfunktion saamiin arvoihin suoraan perustuva menetelmä on intuitiivisesti helposti perusteltavissa, sen varjopuolena on luonnollisen kalibroinnin puuttuminen. Kuinka alas — suhteessa uskottavuusfunktion maksimiarvoon — tulisi kussakin tilanteessa vetää raja alueen sisälle tulevien ja ulkopuolelle jäävien parametriarvojen välille? Koska parametriarvoihin ei voi liittää todennäköisyyksiä, ei myöskään päätöstä kynnysarvon valinnasta voi tehdä niihin vetoamalla.

Voidaan myös todeta, että Bayes-paradigmaa noudattavassa lähestymistavassa tarvetta erilliseen väliestimointiin ja sitä vastaaviin menetelmiin ei — ainakaan periaatteessa — ole, sillä estimointiongelman vastaukseksi saadaan kokonainen estimoitavan parametrin (posteriori)todennäköisyysjakauma. Tämän jakauman avulla puolestaan voidaan määrittää suoraan esim. todennäköisyys sille, että estimoitava parametrin arvo kuuluu jollekin välille — tai yleisemmin johonkin parametriavaruuden osajoukkoon. Samoin, jos parametri on vektorisuure ja siinä on tarkasteltavan ongelman kannalta ylimääräisiä dimensioita, joiden estimoinnista ei olla kiinnostuneita, ne voidaan aina integroida pois posteriorijakaumasta ja siirtyä tarkastelemaan näin syntyvää reunajakaumaa. Estimointiongelman tarkastelu tästä lähtökohdasta edellyttää kuitenkin mallin $\boldsymbol{M}^*$, ts. sekä uskottavuusfunktion että myös priorijakauman määrittämistä, mitä monet tutkijat pitävät Bayes-paradigmaan liittyvänä vakavana puutteena. Voidaan esittää myös vastakkainen argumentti, jonka mukaan henkilön järkevä käyttäytyminen edellyttää nimenomaan tutkittavaa ilmiötä koskevien perusteltujen ennakkokäsitysten huomioon ottamista. Voidaan lisäksi esittää teoreettisia perusteita sille, että ongelmaan liittyvä päätöksentekijän epävarmuus tulisi ilmaista kvantitatiivisesti juuri todennäköisyyskäsitettä käyttäen.

Luottamusväleihin liittyviin tarkasteluihin on luontevinta päätyä jälleen esimerkkien kautta. Tästä syystä käymme läpi edellä kappaleessa 2.6 olevat kolme esimerkkiä uudelleen, mutta nyt väliestimoinnin kannalta.

Odotusarvon luottamusväli riippumattoman otoksen tapauksessa, kun varianssi on tunnettu: Tässä pyrimme siis määrittämään aikaisemmin johdetun (kaava (2.16)) odotusarvon piste-estimaatin $\hat{\mu}(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$, ts. otoskeskiarvon, ympärille sopivan välin, johon voisimme sitten perustellummin ajatella parametrin μ tuntemattoman arvon sijoittuvan. Lähtökohtana on tällöin todennäköisyysarvio

$$P_{(\mu,\sigma^2)}\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha, \tag{3.1}$$

missä $\alpha > 0$ on jokin valittu luku (tavallisimmin 0.05, 0.01 tai 0.001) ja $z_{\alpha/2}$ on se standardi-normaalijakauman (kvantiili)piste, josta oikealle tulevan "häntäalueen" pinta-ala on $\alpha/2$. Jakauman symmetriasta johtuen myös pisteestä $-z_{\alpha/2}$ vasemmalle luetun häntäalueen pinta-ala on $\alpha/2$. Tämä todennäköisyysarvio seuraa suoraan aiemmin johdetusta tiedosta (2.13), jonka mukaan muuttujan $\bar{X}_n$ jakauma on $N(\mu, \sigma^2/n)$, joten siis standardoidun muuttujan $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ jakauma on $N(0,1)$. Muokkaamalla sulkumerkkien sisällä olevaa kaksoisepäyhtälöä arvio (3.1) voidaan edelleen kirjoittaa muotoon, jossa parametri μ on keskellä, ts.

$$P_{(\mu,\sigma^2)}(\bar{X}_n - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n})) = 1 - \alpha. \tag{3.2}$$

Näin voidaan sanoa tuntemattoman μ:n arvon sijaitsevan todennäköisyydellä $1-\alpha$ välillä $\big(\bar{X}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n})\big)$. Jos tämän välin päätepisteet määrittelevään lausekkeeseen nyt sijoitetaan aritmeettisen keskiarvon $\bar{X}_n$ tilalle sen havaittu arvo $\bar{x}_n$, väli saa muodon

$$\big(\bar{x}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x}_n + z_{\alpha/2}(\sigma/\sqrt{n})\big) \tag{3.3}$$

Tätä väliä kutsutaan (havaintoa $\boldsymbol{x}_n$ vastaavaksi) *parametrin μ tason $(1-\alpha)$ luottamusväliksi* (engl. *confidence interval*).

Esimerkki 3.1 *Eräässä koulussa punnittiin* 16 *satunnaisesti valittua poikaoppilasta, joiden painon keskiarvoksi tuli* $\bar{x}_{16} = 52$ *kg. Aiemman kokemuksen perusteella tiedetään, että tämän ikäisten poikien painot vaihtelevat likimäärin normaalijakauman mukaan keskihajontana* 4 *kg. Jakauman odotusarvoa* μ *ei kuitenkaan tunneta ja sitä halutaan arvioida em. otoksen perusteella. Määritä tätä varten* μ*:lle havaintoihin perustuva tason* 0.99 *luottamusväli. (Tenttitehtävä 18.5.2006)*

Ratkaisu. Kun varianssi on tunnettu, saadaan odotusarvon luottamusväliksi kaavan (3.3) mukaisesti

$$\left(\bar{x}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x}_n + z_{\alpha/2}(\sigma/\sqrt{n})\right).$$

Lasketaan $z_{\alpha/2}(\sigma/\sqrt{n}) \approx 2.58 \times (4/\sqrt{16}) = 2.58$, joten odotusarvon μ tason 0.99 luottamusväli on

$$(49.42, 54.58).$$

Pohdintaa. On tärkeää huomata jälleen, että tässä tarkastelussa olemme pysytelleet mallin $\boldsymbol{M}$ puitteissa: Parametri μ on "tuntematon, mutta kiinteä" ja edellä kaavassa (3.2) havaintojen perusteella määräytyvä väli $\left(\bar{X}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n})\right)$ on satunnaismuuttujien funktiona "satunnainen". (Huomaa, että aiemmin tarkastellussa Bayes-päättelyssä näiden muuttujien (kappale 2.7) roolit olivat päinvastaiset: Todennäköisyydet olivat ehdollisia suhteessa havaittuun ja siten "kiinteään" dataan, kun taas tuntematon parametri käsitettiin "satunnaiseksi".) Kun nyt kaavaa (3.3) varten välin

$$\left(\bar{X}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n})\right)$$

lausekkeessa aiemmin satunnaismuuttujiksi tulkittujen otoskeskiarvojen $\bar{X}_n$ tilalle sijoitetaan niiden havaitut arvot $\bar{x}_n$, luonnollisesti myös itse välistä tulee "kiinteä". Tästä syystä tähän *havaittuun luottamusväliin* ei voi sellaisenaan enää liittää mitään todennäköisyysmäärettä, koska sekä itse väli että myös parametri μ ovat "kiinteitä" sen jälkeen kun havainto on jo tehty. Erityisesti olisi siis väärin — vaikkakin houkuttelevaa — sanoa, että "havaittu väli (3.3) sisältää parametrin oikean arvon todennäköisyydellä $1 - \alpha$". Juuri tästä syystä välin (3.3) yhteydessä puhutaankin luottamusvälistä, ei esim. todennäköisyysvälistä.

Joskus, kun halutaan etsiä luottamusväleille käypää frekventististä todennäköisyystulkintaa, päädytään soveltamaan ajatusta, jonka mukaan alkuperäistä koetta toistettaisiin vastaavissa olosuhteissa hyvin monta kertaa peräkkäin. Tätä ilmaisemaan voitaisiin käyttää merkintää "$\boldsymbol{X}_n^{\text{rep}}$", yläindeksin "rep" viitatessa englannin kielen sanaan *repeat(ed)*. Nyt voitaisiin ajatella, että määritettäessä tällaisia havaintoja $\boldsymbol{X}_n^{\text{rep}} = \boldsymbol{x}_n^{\text{rep}}$ vastaavia luottamusvälejä hyvin monta kertaa — jolloin niiden sijainti luonnollisesti vaihtelee kerrasta toiseen havaintojen aritmeettisen keskiarvon mukaan, vaikka μ ja σ pysyisivätkin muuttumattomina — suunnilleen osuus $1-\alpha$ syntyvistä väleistä todella sisältää parametrin μ kun taas osuus α "menee μ:n ohi". Emme kuitenkaan voi koskaan sanoa yksittäisen luottamusvälin kohdalla kumpi näistä vaihtoehdoista on todella tapahtunut.

Varianssin luottamusväli riippumattoman otoksen tapauksessa, kun odotusarvo on tunnettu: Edellä kohdassa 2.6 johdimme varianssin suurimman uskottavuuden estimaattorille lausekkeen $\hat{\sigma}^2(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$ (kaava (2.18)) ja totesimme sitten, että satunnaismuuttuja $n\hat{\sigma}^2(\boldsymbol{X}_n) = \sum_{i=1}^n (X_i - \mu)^2$ jakautuu kuten $\boldsymbol{\chi}^2(n)$-muuttuja

kerrottuna varianssilla σ^2. Näin voimme siis kirjoittaa, edellisen esimerkin tapaan, todennäköisyysarvion

$$P_{(\mu,\sigma^2)}\left(\chi^2_{1-\alpha/2}(n) \leq \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \leq \chi^2_{\alpha/2}(n)\right) = 1-\alpha. \tag{3.4}$$

Tässä $\chi^2_{\alpha/2}(n)$ on se $\boldsymbol{\chi}^2(n)$-jakauman (kvantiili)piste, josta luettuna jakauman oikeanpuoleisen häntäalueen todennäköisyys on $\alpha/2$, ja vastaavasti $\chi^2_{1-\alpha/2}$ se piste, josta oikealle tulevan alueen todennäköisyys on $1-\alpha/2$ (ja siten vasemmalle $\alpha/2$). Muokkaamalla jälleen hieman todennäköisyyslausekkeen (3.4) sisällä olevaa kaksoisepäyhtälöä se voidaan kirjoittaa muotoon

$$P_{(\mu,\sigma^2)}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right) = 1-\alpha.$$

Aivan vastaavalla päättelyllä kuin edellä, sijoittamalla tässä satunnaismuuttujien X_i tilalle niiden havaitut arvot x_i, johdutaan nyt varianssin σ^2 tason $1-\alpha$ luottamusvälin lausekkeeseen

$$\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right)$$

Edellä tehdyt luottamusvälin tulkintaa koskevat huomautukset pätevät tässä lähes sellaisinaan, ainoana erona se, että hypoteettiset $\boldsymbol{X}_n^{\text{rep}}$-havaintoja vastaavat luottamusvälit nyt määritettäisiin tunnuslukujen $\sum_{i=1}^{n}(X_i^{\text{rep}}-\mu)^2$ perusteella.

Odotusarvon luottamusväli riippumattoman otoksen tapauksessa, kun sekä odotusarvo että varianssi ovat tuntemattomia: Tilanteessa, jossa normaalijakauman sekä odotusarvo μ että varianssi σ^2 ovat tuntemattomia parametreja, niille molemmille johdettiin edellä (kaava (2.19)) suurimman uskottavuuden estimaattorin lausekkeet. Periaatteessa voitaisiin siten ajatella, että vain joko μ:n tai σ^2:n luottamusvälin asemesta johdettaisiin molemmille parametreille samanaikaisesti pätevä "luottamusjoukko", ts. jokin sellainen otoksen $\boldsymbol{X}_n$ ja valitun luottamustason $1-\alpha$ perusteella määräytyvä (μ,σ^2)-(puoli)tason $\mathbb{R}\times\mathbb{R}^+$ osajoukko $C(\boldsymbol{X}_n)$, jolle pätee kaikilla μ:n ja σ^2:n arvoilla ehto

$$P_{(\mu,\sigma^2)}\left((\mu,\sigma^2)\in C(\boldsymbol{X}_n)\right) = 1-\alpha.$$

Tässä johdantokurssissa kuitenkin rajoitumme tarkastelemaan vain *odotusarvon μ luottamusväliä*. Tehtävässä on toisaalta otettava huomioon, ettei myöskään varianssin σ^2 "oikeaa arvoa" tunneta, joten sekin täytyy estimoida havainnoista.

Lähtökohta on varsin samanlainen kuin edellä johdettaessa odotusarvon luottamusvälin lauseketta: Tarkastelemme havaintojen aritmeettista keskiarvoa ja pyrimme "standardoimaan sen omalla odotusarvollaan ja keskihajonnallaan". Yhtälön (3.1) antaman mallin mukaan toimittaessa joudutaan kuitenkin toteamaan, että lausekkeessa $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ nimittäjänä oleva (ja luottamusvälin (3.3) pituuden määrittämisen kannalta keskeinen) keskiarvon keskihajonta $\sigma/\sqrt{n}$ on sekin tuntematon. Luonnollinen ajatus tässä tilanteessa on korvata σ estimaattorillaan, joksi tässä on tapana valita kaavan (2.20) mukainen lauseke

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}. \tag{3.5}$$

Näin päädytään tarkastelemaan standardoitua muuttujaa

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}}. \tag{3.6}$$

Seuraavissa todennäköisyystarkasteluissa — paitsi havaintojen aritmeettista keskiarvoa $\bar{X}_n$ — myös kaavan (3.5) määrittelemää keskihajonnan estimaattoria s täytyy tarkastella satunnaismuuttujana. Tilastotieteessä ei kuitenkaan ole tapana tehdä merkinnässä mitään eroa satunnaismuuttujan (3.5) ja sen havaitun arvon välillä, vaan molempia merkitään symbolilla s.

Todennäköisyystarkastelun aluksi täytyy selvittää, miten muuttuja (3.6) jakautuu. Ainekset tähän päättelyyn ovat olemassa jo melkein valmiina: Tiedämme aiemman perusteella, että erotus $\bar{X}_n - \mu$ jakautuu kuten $N(0, \sigma^2/n)$; näin se voidaan esittää muodossa $(\sigma/\sqrt{n})\, Z$, missä Z on standardinormaalimuuttuja. Tarkastellessamme osamäärän $\frac{\bar{X}_n - \mu}{s/\sqrt{n}}$ nimittäjää $s/\sqrt{n}$, voimme pitää lähtökohtana aiempaa tietoa (lause 2.7), jonka mukaan $s^2 = s^2(\boldsymbol{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ jakautuu kuten $\boldsymbol{\chi}^2(n-1)$-muuttuja kerrottuna vakiolla $\sigma^2/(n-1)$. Näin ollen voidaan osamäärä $\frac{\bar{X}_n - \mu}{s/\sqrt{n}}$, kun siitä supistetaan sekä osoittajassa että nimittäjässä esiintyvä tuntematon keskihajontaparametri σ, kirjoittaa muotoon

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \tag{3.7}$$

Tämän muuttujan jakauman selvittämiseksi tarvitaan vielä yhtä lisätietoa, joka sekin sisältyy em. lauseeseen 2.7: kaavan (3.7) osoittajassa ja nimittäjässä olevat otoksen $\boldsymbol{X}_n$ funktioina määritellyt muuttujat ovat tarkastellussa mallissa $\boldsymbol{M}$ riippumattomia. Tällä perusteella yhtälön (3.7) oikean puolen määrittelemän satunnaismuuttujan jakauma kiinnittyykin sitten yksikäsitteisellä tavalla.

Näin määriteltyä satunnaismuuttujaa kutsutaan "*t-jakautuneeksi, vapausastelukuna* $n-1$" ja sitä merkitään tavallisesti symbolilla $t(n-1)$. t-jakaumat ovat standardinormaalijakaumaa laakeampia, varsinkin pienemmillä vapausasteluvuilla, jakauman varianssin kasvaessa sitä mukaa kun vapausasteluku n pienenee. Toisaalta, kun $n \to \infty$, $t(n)$-jakauma lähestyy standardinormaalijakaumaa. t-jakauman oikeanpuoleisia häntätodennäköisyyksiä vastaavat (kvantiili)pisteet on taulukoitu käytännöllisesti katsoen kaikissa tilastotieteen oppikirjoissa. Ne löytyvät myös esimerkiksi verkko-osoitteesta
http://www.statsoft.com/textbook/distribution-tables/.

Jos nyt todennäköisyyttä α vastaavaa $t(n)$-jakauman kvantiilia merkitään $t_\alpha(n)$, saadaan edellisen perusteella todennäköisyysarvio

$$P_{(\mu,\sigma^2)}\left(-t_{\alpha/2}(n-1) \le \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \le t_{\alpha/2}(n-1)\right) = 1 - \alpha. \tag{3.8}$$

Muokkaamalla sulkumerkkien sisällä olevaa lauseketta saadaan edelleen kaavaa (3.2) vastaava tulos

$$P_{(\mu,\sigma^2)}\left(\bar{X}_n - t_{\alpha/2}(n-1)(s/\sqrt{n}) \le \mu \le \bar{X}_n + t_{\alpha/2}(n-1)(s/\sqrt{n})\right) = 1 - \alpha.$$

Näin siis tapauksessa, jossa varianssi σ^2 on tuntematon ja se joudutaan korvaamaan havaittuun otokseen perustuvalla estimaatillaan $s^2 = s^2(\boldsymbol{x}_n)$, päädytään *odotusarvon* μ *luottamusvälin* lausekkeeseen

$$\left(\bar{x}_n - t_{\alpha/2}(n-1)(s/\sqrt{n}), \bar{x}_n + t_{\alpha/2}(n-1)(s/\sqrt{n})\right) \tag{3.9}$$

Huomaa, että tässä sekä keskiarvo $\bar{x}_n$ että luottamusvälin pituuden määrittämisen kannalta tärkeä keskihajontaestimaatti s on laskettu otoksen perusteella. Toisaalta t-jakauman kvantiilipisteiden arvot $t_{\alpha/2}(n-1)$ ovat suurempia kuin standardinormaalijakauman vastaavat pisteet $z_{\alpha/2}$, joten myös niitä vastaavat luottamusvälit ovat keskimäärin pitempiä kuin tapauksessa, jossa varianssi σ^2 olisi tunnettu.

Pohdintaa. Edellä käsitellyt esimerkit koskevat tilanteita, joissa luottamusvälitarkasteluiden pohjana ovat eräät normaalijakauman ominaisuuksiin perustuvat täsmälliset jakaumatulokset. On kuitenkin varsin tavallista, että luottamusvälin määritys perustuu johonkin approksimaatioon, nojautuen lähinnä suurimman uskottavuuden estimaattorien asymptoottista normaalisuutta koskeviin tuloksiin (ks. kappale 2.6). Yksinkertaisen esimerkin tällaisesta keskeiseen raja-arvolauseeseen perustuvasta päättelystä tarjoaa *binomikokeeseen liittyvän parametrin luottamusvälin määritys*: Jos oletamme, että uskottavuusfunktio on muotoa

$$L(\theta) = P_\theta(\boldsymbol{X}_n = \boldsymbol{x}_n) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}, \qquad 0 < \theta < 1,$$

on θ:n suurimman uskottavuuden estimaattori "ykkösten havaittu suhteellinen frekvenssi" $\hat{\theta}(\boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n$ (ks. tulos (1.15)). Toisaalta tämä estimaattori, ymmärrettynä satunnaismuuttujaksi $\hat{\theta}(\boldsymbol{X}_n)$, jakautuu mallin $\boldsymbol{M}$ puitteissa keskeisen raja-arvolauseen perusteella suurilla n:n arvoilla likimäärin normaalisesti odotusarvona θ ja varianssina $\theta(1-\theta)/n$. Koska parametri θ oli tuntematon, on sitä luonnollisesti myös tämän viitejakauman varianssi $\theta(1-\theta)/n$. Toisaalta, koska $\hat{\theta}(\boldsymbol{X}_n) = \bar{X}_n$ on θ:n estimaattorina tarkentuva, voidaan myös varianssia $\theta(1-\theta)/n$ estimoida tarkentuvasti käyttäen lauseketta $\bar{X}_n(1-\bar{X}_n)/n$. Näin voidaan — hieman epämuodollisesti — päätellä, että muuttuja $\frac{\sqrt{n}(\bar{X}_n-\theta)}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}$ noudattaa otoskoon kasvaessa asymptoottisesti $N(0,1)$-jakaumaa. Tästä onkin helppo päätyä, noudattamalla aiempaa mallia, θ:n (likimääräisen) luottamusvälin lausekkeeseen

$$\left(\bar{x}_n - z_{\alpha/2}\sqrt{\bar{x}_n(1-\bar{x}_n)/n}, \bar{x}_n + z_{\alpha/2}\sqrt{\bar{x}_n(1-\bar{x}_n)/n}\right)$$

Huomautus 3.2 *Seuraavissa harjoitustehtävissä tarvittavia tilastollisia taulukoita voi löytää monista oppikirjoista ja verkko-osoitteista, mm. osoitteesta*
http://www.statsoft.com/textbook/distribution-tables/

Harjoitustehtävä 3.1 16*:sta satunnaisesti valitusta savukkeesta mitattiin niiden häkäpitoisuudet. Oletetaan sitten, että mittaustulokset vaihtelevat jonkin verran savukkeesta toiseen, johtuen sekä luonnollisesta vaihtelusta savukkeiden välillä että myös pienistä mittausvirheistä, ja että tätä vaihtelua voidaan tyydyttävästi kuvata normaalijakaumalla. Jakauman keskihajonnan tiedetään olevan* $\sigma = 1.4$ mg*, mutta sen odotusarvoa* μ *ei tunneta. Otoksesta määritettiin pitoisuuksien keskiarvoksi* 9.7 mg*. Määritä tämän perusteella* μ*:n tason* 0.95 *luottamusväli. Miten luottamusväli muuttuu, jos saatu keskiarvo perustuikin otokseen, jonka koko oli* $n = 81$*?*

Harjoitustehtävä 3.2 *Tarkastele edellistä tehtävää mutta nyt olettaen, että myös häkäpitoisuuden keskihajonta* σ *on tuntematon. Oletetaan sitten, että otoksesta on määritetty keskihajonnalle estimaatti* $s = 1.4$ mg*. Määritä tällä perusteella* μ*:n tason* 0.95 *luottamusväli ja vertaa sitä tehtävässä* 3.1 *saatuun tulokseen. Tarkastele jälleen myös tilannetta, jossa otoskoko onkin* $n = 81$*. Kummassa tapauksessa luottamusvälit näissä kahdessa tehtävässä eroavat enemmän toisistaan? Mistä syystä?*

Harjoitustehtävä 3.3 *Tarkastele tehtävän 3.1 mukaista tilannetta, mutta soveltaen siihen nyt Bayes-päättelyä. Oletetaan, että priorikäsitystä μ:n arvosta on kuvattu normaalijakaumalla $N(\mu_0, \sigma_0^2)$, missä $\mu_0 = 10$ ja $\sigma_0^2 = 2$. Määritä tätä vastaava μ:n posteriorijakauma ja vertaa sitä tehtävässä 3.1 määritettyyn luottamusväliin tarkastelemalla (posteriorijakauman suhteen symmetristä) väliä $I_{0.95}$, jolla on ominaisuus*

$$P(\mu \in I_{0.95} \mid \textit{mittaustulokset}) = 0.95.$$

Määritä myös posterioritodennäköisyys sille, että $\mu > 10.5$ mg.

Harjoitustehtävä 3.4 *Erääseen tutkimusraporttiin, joka kertoi lukiolaisten matematiikan taitoja mittaavan testin tuloksista, sisältyi maininta, jonka mukaan "opiskelijoiden keskimääräisen suoritustason 95 prosentin luottamusväli testin mitta-asteikolla oli $(452, 470)$". Voiko tällä perusteella sanoa, että noin 95 prosenttia opiskelijoista sijoittuu matemaattisten taitojensa puolesta välille $(452, 470)$? Perustele vastauksesi.*

Luku 4

Tilastollinen hypoteesintestaus

4.1 Tilastollisen testauksen periaatteista

Tilastollisesta estimoinnista — jota edellä on käsitelty ensin piste-estimoinnin ja sitten väliestimoinnin kannalta — siirrytään nyt tarkastelemaan ns. tilastollista hypoteesintestausta. Usein katsotaan, että perusteltu tieteellinen tietomme ympäröivästä maailmasta karttuu siten, että aluksi esitetään erilaisia mahdollisia tapoja ymmärtää ja jäsentää tarkasteltua ilmiötä täsmällisesti muotoiltujen hypoteesien muodossa ja sen jälkeen pyritään systemaattisia menetelmiä käyttäen osoittamaan tällaiset hypoteesit joko oikeiksi tai vääriksi. Ns. popperilainen perinne korostaa erityisesti jälkimmäistä mahdollisuutta, ts. ajatellaan, että käsitykset tai teoriat ovat ainakin periaatteessa käyttökelpoisia niin kauan kuin ne eivät ole ristiriidassa havaintojen kanssa. Jos ristiriita ilmenee, teoria todetaan kelpaamattomaksi (*falsifioidaan*).

Tutkittavan hypoteesien muotoitu voidaan tavallisesti johtaa suhteellisen suoraan tarkastelun kohteena olevasta ongelmasta. Usein koetetaan selvittää jonkin mahdollisen syyseuraussuhteen olemassaoloa tai sitten arvioida tällaisen riippuvuuden voimakkuutta. Täten voidaan esimerkiksi pyrkiä selvittämään sopivasti valitun kliinisen kokeen avulla, onko tietty uusi syövän hoitomuoto parempi kuin vallitsevan käytännön mukainen hoito. Tätä kysymystä voidaan edelleen täsmentää kysymällä, onko uudella menetelmällä hoidetun potilaan todennäköisyys olla elossa viiden vuoden kuluttua hoidon aloittamisesta suurempi kuin jos hoito olisi tehty käytössä olevalla standardimenetelmällä? Mahdollisia hypoteeseja voisivat tässä tapauksessa olla esimerkiksi "Uusi hoitomuoto johtaa yleensä parempaan lopputulokseen kuin nykyinen", "Nykyinen hoitomuoto johtaa yleensä parempaan lopputulokseen kuin uusi" ja "Hoitomuodot ovat yhtä hyviä". Tämän jälkeen, jotta ongelmaa voitaisiin tarkastella tilastollisen hypoteesintestauksen kannalta, tarvitaan vielä sopiva tilastollinen malli kuvaamaan potilaiden selviytymistä viiden vuoden seurantajakson aikana. Tässä on jälleen taustalla ajatus tilastollisesta mallista dataa generoivana mekanismina.

Kun tilastollinen malli on valittu, tarkastelun kohteena oleva hypoteesi voidaan ymmärtää mallia koskevaksi väitteeksi, jonka paikkansapitävyyttä sitten pyritään arvioimaan havaintojen perusteella. Ahtaammassa mielessä hypoteesi liittyy johonkin valittuun parametriseen malliin, jolloin se tavallisesti sisältää parametrin oikeaa arvoa koskevan, muotoa "$\theta \in A$" olevan väitteen, missä testijoukko A on jokin tarkastellun kysymyksen kannalta sopivalla tavalla valittu parametriavaruuden Θ osajoukko. Väitettä "$\theta \in A$" tarkastellaan tavallisesti suhteessa sen komplementtiin, ts. väitteeseen "$\theta \notin A$". Usein hypoteesi muotoillaan pistemäisenä muodossa "$\theta = \theta_0$" vastaten valintaa $A = \{\theta_0\}$, jolloin vastahypoteesi voidaan kirjoittaa "$\theta \neq \theta_0$". Edellä tarkastellussa hoito-

muotojen vertailuesimerkissä voitaisiin ajatella parametrin vastaavan suoraan elossa olon todennäköisyyttä viiden vuoden kuluttua. Tilanteessa, jossa sekä vanhaa (θ_0) että uutta (θ_1) hoitomuotoa vastaavat parametrinarvot ovat tuntemattomia, johdutaan silloin luonnollisella tavalla vertailemaan vaihtoehtoja "$\theta_0 = \theta_1$", "$\theta_0 > \theta_1$" ja "$\theta_0 < \theta_1$". (Huom. Jos tällaiset vertailut halutaan ilmaista täsmällisesti testijoukkojen avulla, täytyy siirtyä käyttämään pistepareja (θ_0, θ_1) ja valita esim. $A = \{(\theta_0, \theta_1) \in (0,1) \times (0,1) : \theta_0 = \theta_1\}$.

Olemme aiemmin nähneet, että Bayes-paradigman tapauksessa muotoa "$\theta \in A$" olevien väitteiden paikkansapitävyyttä voidaan arvioida suoraan todennäköisyyksien avulla tarkastelemalla niitä parametrinarvoja koskevina tapahtumina $\{\theta \in A\}$. Jos posterioritodennäköisyys $P(\theta \in A | \boldsymbol{X}_n = \boldsymbol{x}_n)$ on lähellä ykköstä, voidaan väitettä "$\theta \in A$" pitää havaintojen $\boldsymbol{X}_n = \boldsymbol{x}_n$ valossa "luultavasti oikeana", jos se taas on lähellä nollaa, "luultavasti vääränä". Bayes-paradigman puitteissa toimittaessa ei näin ollen muodostu selvää rajaa tilastollisen parametriestimoinnin ja hypoteesintestauksen välille.

Frekventistisen paradigman kohdalla tilanne on kuitenkin toinen, koska malliparametreja ei silloin voi tulkita satunnaismuuttujiksi eikä näin ollen myöskään ole mahdollista tarkastella niiden arvoihin liittyviä todennäköisyyksiä. Sen sijaan — kuten edellä on nähty piste- ja väliestimoinnin yhteydessä — frekventistisen paradigman yhteydessä tarkasteltavat todennäköisyydet (tai todennäköisyystiheydet) liitetään havaintoihin tulkitsemalla havainnot satunnaismuuttujiksi X, vastaten havaintojen potentiaalista käyttäytymistä ennen kuin varsinaisia havaintoja on tehty. Toinen mahdollisuus on liittää ne aiemmin kuvatulla tavalla X^{rep}-muuttujiin, joiden arvot "voitaisiin havaita mahdollisesti suoritettavissa myöhemmissä ja alkuperäistä koetta vastaavissa toistokokeissa".

Hyvän lähtökohdan vaihtoehtoisten hypoteesien tarkastelulle saatujen havaintojen valossa tarjoaa asetelma, jossa vertaillaan keskenään eri parametrinarvojen kykyä selittää havaittua dataa. Tällaista selityskykyä on luonnollista mitata suoraan käyttämällä sen mittana havaitun datan uskottavuusfunktion arvoja parametriavaruuden eri pisteissä. Vastaavaa ajatusta sovellettiin aiemmin suurimman uskottavuuden estimointimenetelmän yhteydessä. Yksinkertaisimmillaan tällainen vertailu liittyy kahteen parametrin arvoon, jolloin asetetaan vastakkain esim. hypoteesi $H : \theta = \theta_0$ ja tälle sitten vaihtoehtoinen hypoteesi eli *vastahypoteesi* $A : \theta = \theta_1$. (Tässä ei oteta suoraan kantaa siihen, onko parametriavaruudessa muitakin pisteitä kuin θ_0 ja θ_1, mutta muita vaihtoehtoja ei ainakaan lähtökohtaisesti silloin tarkastella.) Koska vertailu tapahtuu havaintoa $\boldsymbol{X}_n = \boldsymbol{x}_n$ vastaavan uskottavuusfunktion $L(\theta; \boldsymbol{x}_n)$ arvojen avulla, johdumme tarkastelemaan niiden osamäärää

$$LR(\theta_0, \theta_1; \boldsymbol{x}_n) = \frac{L(\theta_1; \boldsymbol{x}_n)}{L(\theta_0; \boldsymbol{x}_n)} \tag{4.1}$$

eli *uskottavuusosamäärää* (engl. *likelihood ratio*).

Esimerkki 4.1 *Olkoon $X_1, X_2, \ldots, X_n$ riippumaton otos* Poisson(μ)*-jakaumasta, missä μ on tuntematon parametri. Tarkastele yksinkertaista hypoteesintestausongelmaa, missä vastakkain ovat hypoteesi $H : \mu = \mu_0$ ja tämän vastahypoteesi $A : \mu = \mu_1$ ($\mu_0 < \mu_1$). Näytä, että tässä ongelmassa uskottavuusosamäärätesti palautuu suoraan muuttujan $\sum_{i=1}^n X_i$ saamien arvojen tarkasteluun. (Tenttitehtävä 25.5.04)*

Ratkaisu. Poisson(μ)-jakautuneen satunnaismuuttujan pistetodennäköisyysfunktio on $p(x; \mu) = e^{-\mu}\frac{\mu^x}{x!}$. Havaintoa $\boldsymbol{X}_n = \boldsymbol{x}_n$ vastaavaksi uskottavuusfunktioksi saadaan pistetodennäköisyyksien tulona

$$L(\mu; \boldsymbol{x}_n) = \prod_{i=1}^n p(x_i; \mu) = \prod_{i=1}^n e^{-\mu}\frac{\mu^{x_i}}{x_i!} = e^{-n\mu}\mu^{\sum_i x_i}\prod_{i=1}^n \frac{1}{x_i!}.$$

Käyttämällä hypoteesien H ja A mukaisia μ:n arvoja uskottavuusosamäärä saa muodon

$$LR(\mu_0, \mu_1; \boldsymbol{x}_n) = \frac{L(\mu_1; \boldsymbol{x}_n)}{L(\mu_0; \boldsymbol{x}_n)} = \frac{e^{-n\mu_1}\mu_1^{\sum_i x_i}\prod_{i=1}^n \frac{1}{X_i!}}{e^{-n\mu_0}\mu_0^{\sum_i x_i}\prod_{i=1}^n \frac{1}{X_i!}} = e^{n(\mu_0-\mu_1)}\left(\frac{\mu_1}{\mu_0}\right)^{\sum_i x_i}.$$

Testisuure riippuu siten satunnaismuuttujien $X_1, X_2, \ldots, X_n$ havaituista arvoista vain niiden summan $\sum_i x_i$ kautta.

Huomautus 4.2 *Jos Bayes-paradigman mukaisessa lähestymistavassa tarkastellaan parametrinarvojen θ_0 ja θ_1 posterioritodennäköisyyksiä (tai jos parametri ymmärretään jatkuvaksi satunnaismuuttujaksi, vastaavia posterioritodennäköisyystiheyksiä), kun ehtona on havaittu data $\boldsymbol{x}_n$, päädytään Bayesin kaavan $p(\theta|\boldsymbol{x}_n) \propto p(\theta)f_X(\boldsymbol{x}_n|\theta) = p(\theta)L(\theta; \boldsymbol{x}_n)$ perusteella heti tulokseen*

$$\frac{p(\theta_1|\boldsymbol{x}_n)}{p(\theta_0|\boldsymbol{x}_n)} = \frac{p(\theta_1)}{p(\theta_0)}\, LR(\theta_0, \theta_1; \boldsymbol{x}_n).$$

Huomaa, että parametrista riippumattomat Bayesin kaavan verrannollisuustekijät $p(\boldsymbol{x}_n)$ supistuvat tässä pois. Näin ollen uskottavuusosamäärä $LR(\theta_0, \theta_1; \boldsymbol{x}_n)$ voidaan ymmärtää tekijäksi, jolla parameterin arvoihin θ_0 ja θ_1 liitettyjen prioritodennäköisyyksien (todennäköisyystiheyksien) suhde $p(\theta_1)/p(\theta_0)$ tulee kertoa, jotta siitä tulisi vastaava posterioritodennäköisyyksien (todennäköisyystiheyksien) suhde.

Testiasetelmassa, jossa vaihtoehtoja $H : \theta = \theta_0$ ja $A : \theta = \theta_1$ tarkastellaan symmetrisinä, tuntuu järkevältä kallistua lähinnä hypoteesin A kannalle silloin kun $LR(\theta_0, \theta_1; \boldsymbol{x}_n) > 1$ ja vastaavasti hypoteesin H kannalle silloin kun $LR(\theta_0, \theta_1; \boldsymbol{x}_n) < 1$. Usein arvo θ_0 on kuitenkin valittu siten, että sitä vastaava hypoteesi H edustaa tietyllä tavalla yksinkertaisempaa selitystä saaduille havainnoille (”ei vaikutusta”, ”ei muutosta aikaisempaan”, tms.) kun taas arvo θ_1 voidaan nähdä pikemminkin edustamassa kokonaista joukkoa tästä yksinkertaisesta ns. *nollahypoteesista* poikkeavia selitysvaihtoehtoja. Jos siis ajatellaan, että yksinkertainen selitys on parempi, voi olla paikallaan pitää kiinni hypoteesista H niin kauan kun se ei ole selvästi ristiriidassa havaintojen kanssa (vrt. edellä mainittu falsifiointiperiaate). H:n kannalta kriittisimpiä ovat luonnollisesti sellaiset havainnot $\boldsymbol{x}_n$, joita vastaava uskottavuusosamäärän $LR(\theta_0, \theta_1; \boldsymbol{x}_n)$ arvo on erityisen suuri, sillä juuri ne selittyvät suhteellisesti ottaen paremmin vaihtoehdon A kuin vaihtoehdon H puitteissa. Tällaista suuriin uskottavuusosamäärän arvoihin johtavaan otosavaruuden osajoukkoa kutsutaan testin *kriittiseksi alueeksi* (engl. *critical region*) eli *hypoteesin H hylkäysalueeksi*. Konkreettisia esimerkkejä kriittisen alueen määrittämisestä esitetään seuraavassa kappaleessa.

4.2 Eräitä tärkeitä normaalijakaumaan perustuvia testejä

Normaalijakauman odotusarvoa koskevan yksinkertaisen hypoteesin testaus riippumattoman otoksen tapauksessa, kun varianssi on tunnettu: Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ on poimittu riippumattomalla otannalla normaalijakaumasta $N(\mu, \sigma^2)$. Oletamme aluksi, että varianssi σ^2 on tunnettu ja tarkastelemme sitten odotusarvoa koskevaa hypoteesia $H : \mu = \mu_0$ ja sille vaihtoehtoista hypoteesia $A : \mu = \mu_1$.

Oletamme tässä, että $\mu_0 < \mu_1$. (Tapaus $\mu_0 > \mu_1$ on muuten samanlainen, mutta siinä johdettujen epäyhtälöiden suunta muuttuu päinvastaiseksi.) Yksinkertaisella laskulla (vrt. kaava (2.15)) voimme todeta, että uskottavuusosamäärän logaritmi on muotoa

$$\begin{aligned}\log LR(\mu_0, \mu_1; \boldsymbol{x}_n) &= \frac{n}{2\sigma^2}\left[(\bar{x}_n - \mu_0)^2 - (\bar{x}_n - \mu_1)^2\right] \\ &= \frac{n}{2\sigma^2}\left[2\left(\mu_1 - \mu_0\right)\bar{x}_n + \left(\mu_1^2 - \mu_0^2\right)\right],\end{aligned}$$

missä $\bar{x}_n$ on havaintojen $x_1, x_2, \ldots, x_n$ aritmeettinen keskiarvo $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Tämä lauseke — ja samoin luonnollisesti myös itse uskottavuusosamäärä $LR(\mu_0, \mu_1; \boldsymbol{x})$ — on $\bar{x}_n$:n kasvava funktio. Uskottavuusosamäärätestin kriittinen alue on siten muotoa

$$R_c = \{\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n) : \bar{x}_n \geq c\}.$$

Tehtäväksi jää näin ollen sopivan kynnysarvon c määrittäminen. Tässä tarkastelussa nousee esiin kaksi erilaista näkökohtaa:

(i) Jos kynnysarvo valitaan kovin matalaksi, havainnot $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ joutuvat hylkäysalueelle suhteellisen helposti myös silloin kun jakauman odotusarvo on nollahypoteesin $H : \mu = \mu_0$ mukainen, jolloin valittu testi siis tuottaa tämän hypoteesin hylkäävän tuloksen "turhan usein". Edellä mainittu varovaisuusperiaate, jonka mukaan tulisi pitäytyä yksinkertaisempaan hypoteesiin H, ellei se ole selvästi ristiriidassa havaintojen kanssa, johtaa nyt muotoa

$$P_{\mu_0}(\boldsymbol{X_n} \in R_c) = P_{\mu_0}(\bar{X}_n \geq c) \leq \alpha$$

olevaan ehtoon. Tässä luku $\alpha > 0$ rajoittaa hylkäysalueeseen joutumisen eli hypoteesin H hylkäämisen todennäköisyyttä siinä tapauksessa, että se itse asiassa pitääkin paikkansa, ts. $\mu = \mu_0$. Jos näin kuitenkin tapahtuu, sanotaan, että kyseessä on *1. lajin virhe* eli *hylkäysvirhe*. Luvusta α käytetään nimitystä *testin koko* tai *merkitsevyystaso* ja sen arvoksi valitaan tavallisimmin joko 0.05 tai 0.01, joskus myös 0.001. Kynnysarvon c ja testin merkitsevyystason α välinen yhteys voidaan nyt todeta seuraavasti: Hypoteesin H pätiessä tiedämme, että satunnaismuuttujan $\bar{X}_n$ jakauma on $N(\mu_0, \sigma^2/n)$. Tästä johdumme ehtoon

$$P(\bar{X}_n > c) = P\left(Z > \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) \leq \alpha,$$

missä Z on standardi-normaalimuuttuja. Kun merkitään jälleen z_α:lla sitä standardinormaalijakauman kvantiilipistettä, josta oikealle jäävän häntäalueen todennäköisyys on α, saadaan tästä kynnysarvolle c ehto $\frac{\sqrt{n}(c-\mu_0)}{\sigma} \geq z_\alpha$, ts.

$$c \geq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}. \tag{4.2}$$

Useimmiten c:n valinta tehdään tarkasti eli asetetaan $c = \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$. Näin ollen, jos havaintoja on paljon ja/tai niiden varianssi on pieni, kynnysarvo c voidaan valita läheltä nollahypoteesin mukaista arvoa μ_0.

(ii) Edellisen perusteella on selvää, ettei kynnysarvon c valinta riipu vaihtoehtoisen hypoteesin $A : \mu = \mu_1$ määrittelevästä parametrinarvosta μ_1. Kynnysarvon c valinnalla on testin toiminnan kannalta olennainen merkitys myös siinä suhteessa, että se

säätelee todennäköisyyttä päätyä vaihtoehtoon A silloin kun se on oikea, ts. malliparametrin arvo on $\mu = \mu_1$. Olisi tietenkin toivottavaa, että tämä todennäköisyys olisi "suuri", ts. lähellä lukua 1. Tätä vastaava vaatimus kirjoitetaan usein muotoon

$$P_{\mu_1}(\boldsymbol{X_n} \in R_c) = P_{\mu_1}(\bar{X}_n \geq c) \geq 1 - \beta, \tag{4.3}$$

missä $\beta > 0$ on jokin valittu (pieni) luku. Käyttämällä aivan vastaavaa päättelyä kuin edellä, tästä päädytään nyt vaatimukseen $\frac{\sqrt{n}(c-\mu_1)}{\sigma} \leq z_{1-\beta}$ eli $\mu_1 - c \geq \frac{z_\beta \sigma}{\sqrt{n}}$. Näin siis, vastaten jokaista valittua lukua $\beta > 0$, ehdon (4.3) täyttyminen riippuu kynnysarvon c ja vastahypoteesin määrittelevän parametrinarvon μ_1 keskinäisestä sijainnista. Toisaalta, aivan kuten edellä, se riippuu havaintojen lukumäärästä n tarkastellussa otoksessa ja niiden varianssista σ^2. Vaikutus on myös samansuuntainen: Jos havaintoja on paljon ja/tai niiden varianssi on pieni, testissä päädytään vaihtoehtoiseen hypoteesiin A suurella todennäköisyydellä silloin kun se on oikea.

Kriittisen alueen todennäköisyyttä malliparametrin μ oikean arvon funktiona kutsutaan testin *voimakkuudeksi*. Tässä voitaisiin puhua myös testin hylkäysvoimasta: Vaihtoehtoisen hypoteesin A pätiessä pitäisi luonnollisesti valita A ja vastaavasti hylätä H. Kun testiasetelma ja testin merkitsevyystaso on valittu, sen voimaa pisteessä $\mu = \mu_1 > c$ voidaan edellisen perusteella lisätä kasvattamalla otoskokoa n. Tarvittavan otoskoon määrittämiseksi asetetaan seuraavaan tyyppinen tavoite:

> Mikäli vaihtoehtoinen hypoteesi on oikea, vähintään tiettyä suuruusluokkaa olevan eron $\mu_1 - \mu_0$ vallitsevan todellisen tilanteen ja nollahypoteesin välillä tulisi käydä ilmi vähintään jollakin annetulla todennäköisyydellä.

Esimerkki 4.3 *Olkoon $X_1, X_2, \ldots, X_n$ riippumaton otos $N(\mu, 1)$-jakaumasta, missä μ on tuntematon parametri. Tarkastele yksinkertaista hypoteesintestausongelmaa, missä vastakkain ovat hypoteesi $H : \mu = 0$ ja tämän vastahypoteesi $A : \mu = \mu_1$ ($\mu_1 > 0$).*

(i) Määritä kynnysarvo c kun $n = 16$ ja testin kooksi valitaan $\alpha = 0.05$.

(ii) Osoita, että n:n pysyessä kiinteänä testin voimakkuus lähestyy ykköstä kun A on oikea ja μ_1:n arvo kasvaa.

(iii) Määritä kynnysarvo c ja otoskoko n siten, että hylkäysvirheen todennäköisyys on korkeintaan 0.05 ja toisaalta testin voimakkuus pisteessä $\mu = \mu_1 = 1$ on vähintään 0.90.

Ratkaisu.

(i) Kun testin koko α on kiinnitetty, tulee kynnysarvon toteuttaa kaavan (4.2) ehto. Tätä ehtoa voidaan havainnollistaa kuvalla 4.1. Sijoittamalla tähän kaavaan $n = 16$ ja $z_{0.05}$ saadaan kynnysarvolle epäyhtälö

$$c \geq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}} = \frac{z_{0.05}}{4} \approx 0.41.$$

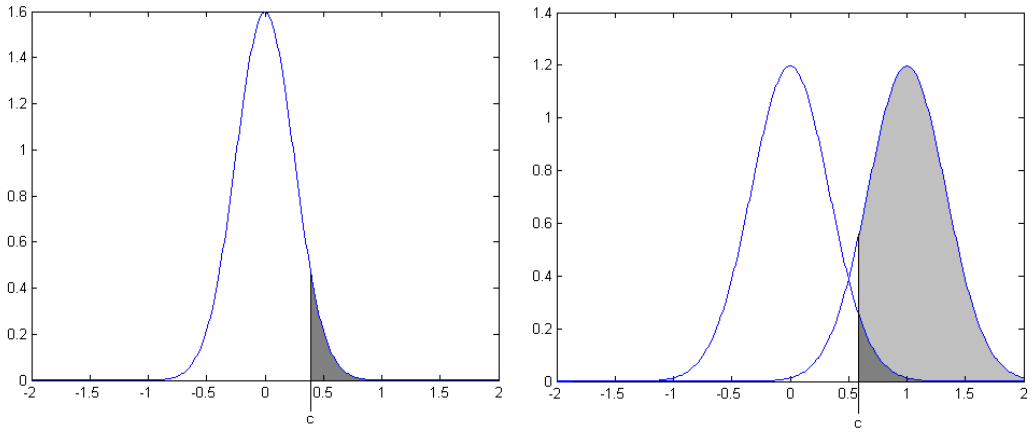
Kynnysarvoksi c tulisi siis asettaa 0.41.

(ii) Testin voima parametrin μ_1 funktiona on

$$\begin{aligned}\pi(\mu_1) &= P_{\mu_1}(\bar{X}_n > c) = P_{\mu_1}\left(\frac{\bar{X}_n - \mu_1}{1/\sqrt{n}} > \frac{c-\mu_1}{1/\sqrt{n}}\right) \\ &= 1 - \Phi((c-\mu_1)\sqrt{n}) = \Phi((\mu_1 - c)\sqrt{n}).\end{aligned}$$

Kun μ_1 kasvaa rajatta, niin $\Phi((\mu_1 - c)\sqrt{n})$ — ja siten myös testin voima — lähestyy ykköstä.

(iii) Tilanne on havainnollistettu kuvassa 4.1. Alaraja kynnysarvolle c saadaan saadaan samalla tavalla kuin kohdassa (i), merkiten kuitenkin otoskokoa (jota ei nyt ole kiinnitetty) jälleen n:llä. Silloin saadaan kynnysarvon alarajaksi $c \geq \frac{z_{0.05}}{\sqrt{n}} \approx \frac{1.65}{\sqrt{n}}$. Testin voima voidaan laskea kuten kohdassa (ii). Asettamalla testin voiman pisteessä



Kuva 4.1: Vasemmanpuoleisessa kuvassa esimerkin 4.3 kohdan (i) tilanne havainnollistettuna. Kynnysarvo c tulisi valita siten, että tummennetun alueen pinta-ala olisi korkeintaan 0.05. Oikeanpuoleisessa kuvassa on saman tehtävän kohdan (iii) tilanne yksinkertaistettuna. Tässä kynnysarvo c tulisi valita siten, että tummemmalla varjostetun alueen pinta ala olisi korkeintaan 0.05 ja molempien varjostettujen alueiden pinta-ala yhteensä vähintään 0.90. Kohdassa (iii) tulee myös määrätä otoskoko n, joka vaikuttaa kuvaajien muotoon ja siten varjostettujen alueiden pinta-aloihin.

$\mu = \mu_1 = 1$ vähintään 0.9:ksi saamme epäyhtälön

$$\pi(1) = \Phi((1-c)\sqrt{n}) \geq 0.9.$$

Tämä epäyhtälö voidaan ratkaista c:n suhteen, jolloin saadaan kynnysarvon ylärajaksi $c \leq 1 - \frac{1.28}{\sqrt{n}}$. Pienin molemmat ehdot täyttävä otoskoko saadaan nyt ratkaisemalla yhtälö

$$1 - \frac{1.28}{\sqrt{n}} = \frac{1.65}{\sqrt{n}} \iff \sqrt{n} = 2.93.$$

Valitsemalla otoskooksi 9 ja kynnysarvoksi $c = \frac{1.65}{3} = 0.55$ vaaditut ehdot täyttyvät.

Pohdintaa. Tapauksessa, jossa havainto ei osu kriittisen alueen sisälle, sanotaan tavallisesti, että "hypoteeesi H jää voimaan". Tällä sanonnalla halutaan korostaa sitä, että vaikka tilastollisella testillä tietyssä mielessä on kyky tai voima johtaa paikkansapitämättömän hypoteesin hylkäykseen, sillä ei voi "todistaa" hypoteesia oikeaksi. Jos hypoteesille H on olemassa hyviä perusteita, sen voidaan ajatella pysyvän voimassa tai ainakin mahdollisena selityksenä havainnoille niin kauan, kun sitä ei ole osoitettu vääräksi (falsifioitu). Joskus kuitenkin — vaikkakaan se ei ole suositeltavaa — puhutaan myös asetetun hypoteesin H hyväksymisestä. Jos tämä tapahtuu tilanteessa, jossa vaihtoehtoinen hypoteesi A olisikin oikea, sanotaan tapahtuvan *2. lajin virhe* eli *hyväksymisvirhe*. Jos edellä oleva ehto (4.3) nyt kirjoitetaan muotoon $P_{\mu_1}(\boldsymbol{X}_n \notin R_c) \leq \beta$, voimme tulkita sen rajoittavan juuri hyväksymisvirheen todennäköisyyttä.

Edellä todettiin myös, ettei valittu kriittinen alue riipu vaihtoehtoisen hypoteesin A määrittelevästä parametrinarvosta μ_1. Näin ollen voidaan ajatella, että testi pysyy samana, vaikka hypoteesin $H : \mu = \mu_0$ vastahypoteesiksi valittaisiinkin yleisempi väite $A : \mu > \mu_0$. Jos parametriavaruudeksi valitaan puoliakseli $\Theta = [0, \infty)$, voidaan kirjoittaa $H : \mu \in \{\mu_0\}$ ja $A : \mu \in (\mu_0, \infty)$. Tällaisessa tilanteessa, jossa hypoteesin määrittelemä parametriavaruuden osa sisältää vain yhden pisteen ja vastahypoteesin määrittelemä osa koostuu useammasta kuin yhdestä pisteestä, sanotaan, että hypoteesi H on *yksinkertainen* ja vastahypoteesi A *yhdistetty*.

Huomautus 4.4 *Mikäli tämä hypoteesintestausongelma haluttaisiin ratkaista Bayes-päättelyn avulla, siinä päädyttäisiin lopuksi tarkastelemaan parametriavaruuden osajoukkojen $\{\mu_0\}$ ja (μ_0, ∞) posterioritodennäköisyyksiä. Jos prioritodennäköisyydet määriteltäisiin jatkuvan jakauman avulla, olisi myös posteriorijakauma jatkuva. Siinä tapauksessa yhteen pisteeseen $\{\mu_0\}$ liitettäisiin aina vain todennäköisyys 0, joten hypoteesi H tulisi aina hylätyksi ja vaihtoehtoinen hypoteesi A vastaavasti aina hyväksytyksi. Tästä syystä, mikäli tätä testiongelmaa halutaan käsitellä bayesläisestä näkökulmasta epätriviaalilla tavalla, täytyy pisteeseen $\{\mu_0\}$ jo alun perin liittää positiivinen prioritodennäköisyys.*

Monesti uskottavuusosamäärän voidaan osoittaa olevan jonkun havainnoista määräytyvän tilastollisen tunnusluvun monotoninen (esim. kasvava) funktio. Tämä ns. MLR-ominaisuus (*monotone likelihood ratio*) yleistää edellisen esimerkin tilanteen, jossa uskottavuusosamäärän riippuvuus otoksesta tapahtui havaintojen summan (tai niiden keskiarvon) kautta ja oli sen kasvava funktio. Useimmiten tilastollisten testien konstruktiossa kuitenkin nojaudutaan suoraan johonkin otoksesta laskettavaan tunnuslukuun, jonka ajatellaan ilmaisevan tai mittaavan mahdollista poikkeamaa hypoteesissa ilmaistusta mallioletuksesta. Vastaavasti tunnusluvun valinta kohdistuu usein suoraan johonkin testin kohteena olevan malliparametrin estimaattoriin. Tilastollisen testin ominaisuuksien toteamisen kannalta tarpeelliset todennäköisyystarkastelut tehdään sitten ko. estimaattorin (otos)jakaumaan perusteella. Tarkastelemme esimerkkinä seuraavaa varianssia koskevaa testiongelmaa:

Normaalijakauman varianssia koskevan yksinkertaisen hypoteesin testaus riippumattoman otoksen tapauksessa: Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ on poimittu riippumattomasti normaalijakaumasta $N(\mu, \sigma^2)$. Jos odotusarvo μ on tunnettu, voimme tarkastella varianssin σ^2 suurimman uskottavuuden estimaattoria

$$\hat{\sigma}^2(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2. \tag{4.4}$$

Aiemman perusteella tiedämme, että oikean puolen neliösumma $\sum_{i=1}^{n}(X_i-\mu)^2$ jakautuu kuten $\chi^2(n)$-muuttuja kerrottuna vakiolla σ^2. Jos nyt asetetaan hypoteesi $H:\sigma^2=\sigma_0^2$ ja tälle vaihtoehtoinen hypoteesi $A:\sigma^2=\sigma_1^2$, missä $\sigma_1^2>\sigma_0^2$, voidaan aluksi rajoittua muotoa $R_c=\{\boldsymbol{x}=(x_1,x_2,...,x_n):\sum_{i=1}^{n}(x_i-\mu)^2\geq c\}$ oleviin kriittisiin alueisiin. Valittua merkitsevyystasoa $\alpha>0$ vastaava vakion c arvo saadaan puolestaan siitä, että hypoteesin H pätiessä neliösumma $\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2$ jakautuu kuten $\chi^2(n)$-muuttuja. Toisaalta, jos havainnot on poimittu jakaumasta, jonka varianssi σ_1^2 on suurempi kuin σ_0^2, tämä tyypillisesti heijastuu myös neliösumman $\sum_{i=1}^{n}(X_i-\mu)^2$ käyttäytymisessä ja sen todennäköisyys joutua valitulle kriittiselle alueelle tulee siten α:aa suuremmaksi. Tässäkään esimerkissä kriittinen alue ei riipu siitä mikä arvo vaihtoehtoisen hypoteesin varianssilla σ_1^2 on, kunhan vain järjestys $\sigma_1^2>\sigma_0^2$ säilyy. Testin voimakkuus sen sijaan kasvaa varianssin σ_1^2 mukana.

Tämän esimerkin mukainen testi muuttuu vain hieman, jos odotusarvoa μ ei tunneta vaan se joudutaan estimoimaan otoksesta. Tällöin johdumme aluksi ennestään tuttuun harhattoman varianssiestimaattorin lausekkeeseen (kaava (2.20))

$$s^2=s^2(\boldsymbol{X}_n)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2, \tag{4.5}$$

josta tiedämme edelleen, että neliösumma $\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$ jakautuu kuten $\chi^2(n-1)$-muuttuja kerrottuna vakiolla σ^2. Muilta osin edellä esitetty testin konstruktio ja sitä koskevat johtopäätökset säilyvätkin voimassa sellaisinaan.

4.3 t-testi

Kuten edellä todettiin, useimmiten tilastollisten testien konstruktiossa nojaudutaan suoraan johonkin otoksesta laskettavaan tunnuslukuun, jonka ajatellaan mittaavan malliparametrin oikean arvon mahdollista poikkeamaa hypoteesissa ilmaistusta oletuksesta. Vastaavasti tunnusluvun valinta kohdistuu usein suoraan johonkin testin kohteena olevan malliparametrin estimaattoriin. Tilastollisen testin ominaisuuksien toteamisen kannalta tarpeelliset todennäköisyystarkastelut tehdään sitten ko. estimaattorin (otos)jakauman perusteella. Tarkastelemme nyt esimerkkinä tällaisesta menettelystä seuraavaa odotusarvoa koskevaa testiongelmaa:

Normaalijakauman odotusarvoa koskevan yksinkertaisen hypoteesin testaus riippumattoman otoksen tapauksessa, kun varianssia ei tunneta (t-testi): Tutkimme nyt miten odotusarvoa μ koskeva hypoteesintestausongelma muuttuu aiemmin käsiteltyyn normaalijakaumatestiin verrattuna silloin, kun odotusarvon lisäksi myös varianssi σ^2 on tuntematon parametri. Johtaessamme edellä vastaavassa tilanteessa odotusarvon luottamusväliä, päädyimme aluksi tarkastelemaan tunnuslukua (vrt. kaava (3.7))

$$t_{n-1}(\boldsymbol{X}_n)=\frac{\bar{X}_n-\mu}{s(\boldsymbol{X}_n)/\sqrt{n}}, \tag{4.6}$$

missä vasemmalla puolella käytetty merkintä viittaa suoraan tämän muuttujan jakaumaan, ts. t-jakaumaan vapausastelukuna $n-1$. Muuttujan (4.6) voidaan katsoa mittaavan aritmeettisen keskiarvon poikkeamaa odotusarvosta, kun tälle poikkeamalle käytetään mittayksikkönä otoksesta estimoitua keskiarvon keskihajontaa $s(\boldsymbol{X}_n)/\sqrt{n}$. Huomaa myös

se tärkeä ominaisuus, ettei muuttujan (4.6) jakauma riipu tuntemattomasta varianssista σ^2, joten varianssi voidaan unohtaa haettaessa sopivaa testimenettelyä odotusarvoa koskevalle hypoteesille.

Noudatamme nyt edellä esitettyä tapaa määritellä odotusarvoa koskeva yksinkertainen hypoteesi ja tälle vastahypoteesi, esittäen hypoteesi väitteenä $H : \mu = \mu_0$ ja vastahypoteesi väitteenä $A : \mu = \mu_1$ ($\mu_0 < \mu_1$). Kun testi konstruoidaan suoraan muuttujan (4.6) perusteella, on luonnollista valita sen hylkäysalue tämän testimuuttujan saamien suurien arvojen perusteella, ts. tarkastelemalla sellaisia mahdollisia otoksia $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$, joille pätee muotoa $t_{n-1}(\boldsymbol{x}_n) \geq c$ oleva ehto. Tämän jälkeen testin konkreettista määrittelyä varten täytyy vain valita sopiva kynnysarvo c. Menetelmä vastaa aiemmin esitetyn kaavan (3.8) vaatimusta, mutta nyt valittua testin kokoa vastaava ehto saa muodon

$$P_{(\mu_0,\sigma^2)}\left(\frac{\bar{X}_n - \mu_0}{s(\boldsymbol{X}_n)/\sqrt{n}} \geq c\right) = P_{(\mu_0,\sigma^2)}\left(t_{n-1}(\boldsymbol{X}_n) \geq c\right) \leq \alpha,$$

missä testimuuttujan $t_{n-1}(\boldsymbol{X}_n)$ määritelmässä (4.6) hypoteesin H mukaisesti asetetaan $\mu = \mu_0$. Muuttujan $t_{n-1}(\boldsymbol{X}_n)$ jakaumaa koskevan tuloksen perusteella voidaan kynnysarvoksi valita $c = t_\alpha(n-1)$, ts. se $t(n-1)$-jakauman (kvantiili)piste, jonka oikealle puolelle jäävän häntäalueen pinta-ala on α. Tästä päädytään tilastolliseen testiin, jossa hypoteesi H hylätään, mikäli havaittu testimuuttujan arvo $t_{n-1}(\boldsymbol{x}_n)$ ylittää kynnysarvon c, ts. mikäli

$$t_{n-1}(\boldsymbol{x}_n) = \frac{\bar{x}_n - \mu_0}{s(\boldsymbol{x}_n)/\sqrt{n}} > t_\alpha(n-1).$$

Tämä ns. t-testi vastaa läheisesti edellä normaalijakaumatestin tapauksessa johdettua tulosta. Huomaa kuitenkin, että tässä myös hajontaestimaatti s on otoksen $\boldsymbol{x}_n$ funktio, vaikka tätä seikkaa ei aihetta käsittelevässä kirjallisuudessa tavallisesti korostetakaan merkinnällä $s = s(\boldsymbol{x}_n)$. Hypoteesin H mahdollinen hylkääminen ei siten tässä tapauksessa riipu pelkästään otoksesta määritetystä aritmeettisesta keskiarvosta $\bar{x}_n$, vaan myös sen perusteella lasketusta hajontaestimaatista $s(\boldsymbol{x}_n)$.

Kuten aiemmin todettiin, $t(n)$-jakaumat muistuttavat muodoltaan normaalijakaumaa, mutta ne ovat sitä laakeampia, etenkin pienemmillä n:n arvoilla. Täten t-testi johtaa keskimäärin harvemmin odotusarvoa koskevan hypoteesin H hylkäämiseen kuin vastaava normaalijakaumatesti, kun verrataan keskenään tapauksia, joissa hypoteesin mukaisen odotusarvon μ_0 ja todellisen odotusarvon välinen erotus on sama mutta joissa havaintojen varianssi on edellisessä tapauksessa tuntematon ja jälkimmäisessä tunnettu. Toisin ilmaistuna, odotusarvoa koskevan t-testin voimakkuus parametrin μ funktiona on pienempi kuin vastaavan normaalijakaumatestin. Tämä on luonnollista, kun otetaan huomioon t-testin tapauksessa ilmenevä tarve estimoida samalla myös jakauman varianssia, jolloin päättelyyn liittyy enemmän epävarmuutta. Vielä on syytä huomata, että myös t-testi on riippumaton siitä, millä tavalla vastahypoteesi $A : \mu = \mu_1$ asetetaan, kunhan vain $\mu_0 < \mu_1$. Testi säilyy siten muuttumattomana, jos vastahypoteesina tarkastellaan yhdistettyä hypoteesia $A : \mu > \mu_0$.

4.4 Yksi- ja kaksisuuntaiset testit ja eräitä muita tarkasteluja

Joskus odotusarvoa koskevan hypoteesin $H : \mu = \mu_0$ vastahypoteesia on paikallaan tarkastella kaksisuuntaisena, ts. ottaen H:n vaihtoehtoina huomioon sekä mahdollisuus

$\mu > \mu_0$ että mahdollisuus $\mu < \mu_0$. Näin on syytä menetellä etenkin tapauksissa, joissa tavoitteena on jonkin väitetyn syy-seuraussuhteen osoittaminen tai kumoaminen sopivalla koejärjestelyllä kerättyjen havaintojen avulla, mutta mahdollisen vaikutuksen olemassaolon ohella myös sen suunta on avoin kysymys. Tällöin H:n vastahypoteesiksi voidaan asettaa väite $A : \mu \neq \mu_0$.

Jos tarkastellaan normaalijakaumaa, jossa varianssi on tunnettu, niin päädytään hylkäämään hypoteesi H valittua merkitsevyystasoa α vastaavassa testissä, jos otoksesta määritetty havaintojen aritmeettinen keskiarvo toteuttaa jomman kumman ehdoista

$$\bar{x}_n < \mu_0 - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad \text{tai} \quad \bar{x}_n > \mu_0 + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Nämä ehdot voi tietenkin myös yhdistää kirjoittamalla

$$|\bar{x}_n - \mu_0| > \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Tämä epäyhtälö ilmaisee sen luonnollisen periaatteen, että "hypoteesin H kannalta kriittisiä ovat sellaiset $\bar{x}_n$:n arvot, jotka sijaitsevat kaukana sen määrittelevästä arvosta μ_0". Huomaa myös, että päinvastaisessa tapauksessa, ts. kun

$$\mu_0 - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \bar{x}_n < \mu_0 + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \tag{4.7}$$

hypoteesi H "jää voimaan". Samalla huomataan, että kaksoisepäyhtälö (4.7) lausuu täsmälleen saman ehdon kuin se, jonka avulla aiemmin määrittelimme odotusarvon μ tason $(1-\alpha)$ luottamusvälin lausekkeen $\big(\bar{x}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x}_n + z_{\alpha/2}(\sigma/\sqrt{n})\big)$ (ks. kaava (3.3)). Hypoteesi $H : \mu = \mu_0$ jää voimaan merkitsevyystason α testissä täsmälleen silloin kun sen määrittelemä parametrinarvo $\mu = \mu_0$ kuuluu μ:n tason $1-\alpha$ luottamusvälin sisälle. Aivan vastaavanalainen yhteys vallitsee kaksisuuntaista vaihtoehtoista hypoteesia $A : \mu \neq \mu_0$ vastaavan t-testin hyväksymisalueen ja odotusarvolle johdetun luottamusvälin lausekkeen (3.9) välillä tarkasteltaessa testi- ja estimointiongelmia tilanteessa, jossa varianssia ei tunneta.

Kahden odotusarvon vertailu normaalijakautuneiden havaintojen tapauksessa: Käytännössä varsin usein — kuten edellä lyhyesti tarkastellussa esimerkissä, jossa ajateltiin verrattavan toisiinsa vanhaa ja uutta hoitomenetelmää — päädytään testiongelmaan, jossa hypoteesi ei sisällä varsinaista väitettä tietyn, tarkastelun kohteena olevan parametrin arvosta vaan se koskee kahden eri parametrin "oikeiden arvojen" mahdollista yhtäsuuruutta. Tutkimme ensin testiongelmaa, jossa vertaillaan kahden normaalijakautuneen osaotoksen odotusarvoja tilanteessa, jossa näiden otosten varianssit tunnetaan. Olkoon siis $X_{11}, X_{12}, \ldots, X_{1n_1}$ riippumaton otos $N(\mu_1, \sigma_1^2)$-jakaumasta ja $X_{21}, X_{22}, \ldots X_{2n_2}$ vastaavasti ensimmäisestä otoksesta riippumaton otos $N(\mu_2, \sigma_2^2)$-jakaumasta. Asetamme nyt hypoteesin $H : \mu_1 = \mu_2$ ja tälle vastahypoteesin $A : \mu_1 \neq \mu_2$. Tässä voimme soveltaa kummankin otoksen kohdalla aiemmin johdettua tulosta, jonka mukaan osakeskiarvo $\bar{X}_{1n_1}$ noudattaa normaalijakaumaa $N(\mu_1, \sigma_1^2/n_1)$ ja osakeskiarvo $\bar{X}_{2n_2}$ normaalijakaumaa $N(\mu_2, \sigma_2^2/n_2)$. Koska ne ovat oletuksen mukaan myös riippumattomia satunnaismuuttujia, noudattaa niiden erotus $\bar{X}_{1n_1} - \bar{X}_{2n_2}$ normaalijakaumaa $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. Hypoteesin H pätiessä tämän jakauman odotusarvo on 0, joten voimme kytkeä tämän testiongelman suoraan aiemmin käsiteltyyn normaalijakaumatestiin $H : \mu = 0$, $A : \mu \neq 0$ asettamalla siinä $\mu = \mu_1 - \mu_2$ ja $\sigma^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$.

Siinä tapauksessa, että varianssien σ_1^2 ja σ_2^2 arvoja ei tunneta, mutta ne kuitenkin oletetaan samansuuruisiksi, odotusarvojen vertailutehtävä hypoteesintestausmenetelmällä palautuu helposti jo edellä käsiteltyyn t-testiin. Tämä nähdään seuraavasti: Aiemmin on todettu, että otoksen perusteella laskettu poikkeusneliösumma aritmeettisen keskiarvon suhteen jakautuu kuten $\chi^2(n-1)$-muuttuja, missä n on havaintojen lukumäärä, kerrottuna havaintojen varianssilla σ^2. Tätä tulosta voidaan nyt soveltaa kumpaankin osaotokseen, jolloin saadaan tulokset

$$\sum_{i=1}^{n}(X_{1i}-\bar{X}_{1n_1})^2 \sim \sigma^2\chi^2(n_1-1) \quad \text{ja} \quad \sum_{i=1}^{n}(X_{2i}-\bar{X}_{2n_2})^2 \sim \sigma^2\chi^2(n_2-1),$$

jossa neliösummat ovat riippumattomia. Koska χ^2-jakauma säilyy riippumattomien χ^2-muuttujien yhteenlaskussa — niiden vapausasteet lasketaan silloin yhteen — saamme tästä jakaumatuloksen

$$\sum_{i=1}^{n}(X_{1i}-\bar{X}_{1n_1})^2 + \sum_{i=1}^{n}(X_{2i}-\bar{X}_{2n_2})^2 \sim \sigma^2\chi^2(n_1+n_2-2).$$

Kun tämä yhdistetään edellä saatuun osakeskiarvojen erotusta koskevaan tulokseen, jonka mukaan erotus $\bar{X}_{1n_1}-\bar{X}_{2n_2} \sim N(\mu_1-\mu_2, \sigma^2(1/n_1+1/n_2))$, päädytään tässä tarkastelemaan testimuuttujana osamäärää

$$t_{n_1+n_2-2}(\boldsymbol{X}_{n_1}, \boldsymbol{X}_{n_2}) = \frac{\bar{X}_{1n_1}-\bar{X}_{2n_2}}{\sqrt{\dfrac{\sum_{i=1}^{n}(X_{1i}-\bar{X}_{1n_1})^2+\sum_{i=1}^{n}(X_{2i}-\bar{X}_{2n_2})^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \qquad (4.8)$$

joka noudattaa hypoteesin $H:\mu_1=\mu_2$ pätiessä $t(n_1+n_2-2)$-jakaumaa. Vastaavassa testissä satunnaismuuttujien tilalle sijoitetaan jälleen niiden havaitut arvot ja verrataan tulosta $t(n_1+n_2-2)$-jakauman kvantiilipisteiden perusteella valittuihin kynnysarvoihin.

Pohdintaa: merkitsevyystestit ja p-arvot. Edellä tarkasteltujen tilastollisten testien konstruktiot perustuivat etukäteen valittua merkitsevyystasoa $\alpha > 0$ vastaavan hylkäysalueen muodostamiseen, jolloin testin antama tulos sitten riippui siitä, sijoittuiko tehty havainto tämän hylkäysalueen sisä- vai ulkopuolelle. Hylkäysalueet on järkevää valita sisäkkäisiksi siten, että pienempää α:n arvoa vastaava alue, joka vastaa konservatiivisempaa suhtautumista H:n hylkäämiseen, aina sisältyy suurempaa α:n arvoa vastaavaan hylkäysalueeseen. Tällöin saattaa hyvinkin syntyä tilanne, jossa esimerkiksi testattava hypoteesi H jää voimaan valitun tason $\alpha = 0.01$ testissä, mutta olisi toisaalta tullut hylätyksi tason $\alpha = 0.05$ testissä. Silloin tulee mieleen mahdollisuus hakea näiden kahden tason väliltä juuri se merkitsevyystaso, joka saatujen havaintojen valossa muodostaisi eräänlaisen vedenjakajan hypoteesin H ja sen vastahypoteesin A valinnan välillä. Tällaista lukua kutsutaan *p-arvoksi*. Usein tilastollisen testin tulos ilmaistaankin käytännössä viittaamalla suoraan testimenettelyn antamaan p-arvoon. Tässä taustana on ajatus, että kuta pienempi saatu p-arvo on, sitä "tilastollisesti merkitsevämpi" on ero saatujen havaintojen ja hypoteesin H välillä. Käytännössä p-arvoja usein pidetäänkin eräänlaisena metriikkana havaintojen ja hypoteesin välillä (tosin käänteisenä siinä mielessä, että pienempi p-arvo viittaa suurempaan etäisyyteen).

Tarkastelemme nyt lyhyesti p-arvon määritystä edellä käsitellyn odotusarvoon liittyvän yksisuuntaisen t-testin yhteydessä. Nämä tarkastelut voidaan helposti siirtää koskemaan muitakin vastaavia testitilanteita. Kuten edellä todettiin, noudattaa testimuuttuja $t_{n-1}(\boldsymbol{X}_n) = \frac{\bar{X}_n - \mu}{s(\boldsymbol{X}_n)/\sqrt{n}}$ hypoteesin $H : \mu = \mu_0$ pätiessä $t(n-1)$-jakaumaa. Jos nyt on tehty havainto $\boldsymbol{X}_n = \boldsymbol{x}_n$, voidaan aluksi laskea sitä vastaava testimuuttujan arvo $t_{n-1}(\boldsymbol{x}_n) = \frac{\bar{x}_n - \mu}{s(\boldsymbol{x}_n)/\sqrt{n}}$ ja määrittää sitten tätä pistettä vastaava $t(n-1)$-jakauman häntätodennäköisyys

$$P_{(\mu_0,\sigma^2)}\left(t_{n-1}(\boldsymbol{X}_n^{\text{rep}}) > t_{n-1}(\boldsymbol{x}_n)\right).$$

Näin saatua lukua kutsutaan *havaintoa $\boldsymbol{x}_n$ vastaavaksi p-arvoksi* ko. yksisuuntaisessa testissä. Tässä on noudatettu aiempaa käytäntöä, jossa "mahdollisesti myöhemmin, alkuperäistä koetta vastaavissa olosuhteissa, tehtävää havaintoa" merkitään symbolilla $\boldsymbol{X}_n^{\text{rep}}$. (Joskus kontrastia tämän satunnaismuuttujan ja todella tehtyjen havaintojen eli datan $\boldsymbol{x}_n$ välillä korostetaan vielä kirjoittamalla jälkimmäisen tilalla selvyyden vuoksi $\boldsymbol{x}_n^{\text{obs}}$). p-arvon määritelmä voitaisiin siten tässä tapauksessa ilmaista seuraavasti: se on *todennäköisyys saada hypoteesin H pätiessä testimuuttujan arvo, joka on sen todellisuudessa havaittua arvoa suurempi.* Saatua havaintoa vastaava p-arvo voidaan määrittää suoraan testimuuttujan saaman arvon $t_{n-1}(\boldsymbol{x}_n^{\text{obs}})$ ja $t(n-1)$-jakauman taulukoiden avulla.

On kuitenkin tärkeää todeta, että p-arvoa **ei saa** tulkita "hypoteesin H todennäköisyydeksi kun ehtona on havaittu data $\boldsymbol{x}_n^{\text{obs}}$". Tällainen väärinkäsitys on varsin yleinen, sillä siihen houkuttelevat tilastollisten testien käyttöön johtavat konkreettiset motiivit. Se vastaa suunnilleen aiemmin luottamusvälien yhteydessä mainittua virhepäätelmää, jonka mukaan "parametrin oikea arvo sijaitsee tason $(1-\alpha)$ luottamusvälin sisällä todennäköisyydellä $1-\alpha$". Perimmältään kummassakin on kysymys siitä, ettei tilastotieteen frekventististä paradigmaa noudattavassa päättelyssä saa liittää todennäköisyyksiä parametrinarvoihin. Bayes-paradigman mukaan toimittaessa tätä vastaavaa rajoitusta ei ole, joten sen puitteissa on täysin mahdollista puhua mm. hypoteesin todennäköisyydestä.

On myös hyvä huomata, että p-arvo riippuu, paitsi havaitusta erosta esim. otoskeskiarvon ja hypoteesin mukaisen odotusarvon välillä, myös käytettävissä olevien havaintojen lukumäärästä. Tästä syystä pienikin määrällinen ero hypoteesin mukaisen teorian ja havaintojen asiasta tuottaman informaation välillä voidaan usein osoittaa "tilastollisesti merkitseväksi" kunhan havaintoja vain kerätään riittävän paljon. Onkin tärkeää, ettei esimerkiksi tieteellisessä raportoinnissa sekoiteta toisiinsa käsitteitä (*määrällisesti* ja/tai *tieteellisesti*) *merkittävä* ja (*tilastollisesti*) *merkitsevä.* (Huom. Suomessa eroa näiden käsitteiden välillä voidaan helposti korostaa käyttämällä vastaavissa yhteyksissä sanoja "merkittävä" ja "merkitsevä". Sen sijaan englannin kielessä vakiintuneessa käytössä on vain sana *significant* riippumatta siitä, kumpi täsmällinen merkitys sille annetaan. Jos siis kysymys on tilastollisesti merkitsevästä erosta hypoteesin ja havaintojen välillä, on tätä englanninkielisessä testissä syytä korostaa käyttämällä termiä *statistically significant.*)

Esimerkki 4.5 *Eräässä tutkimuksessa pyrittiin selvittämään liikunnan vaikutusta veren kolesterolin triglyserolipitoisuuteen sokeritautia sairastavilla tytöillä. Tutkimukseen valittiin 11 tyttöä, jotka viettivät viikon pituisen hiihtoleirin Ruotsin tuntureilla. Kolesterolin triglyserolipitoisuus mitattiin kultakin tytöltä kaksi päivää ennen leirin alkua (mittaustulos X_1) ja viisi päivää sen loppumisen jälkeen (mittaustulos X_2). Mittaustulokset olivat seuraavat:*

Tyttö	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
X_1	*72*	*60*	*102*	*54*	*57*	*42*	*45*	*56*	*61*	*64*	*51*
X_2	*107*	*86*	*123*	*102*	*70*	*70*	*73*	*61*	*100*	*41*	*68*

Tukevatko havainnot hypoteesia, jonka mukaan liikunta vaikuttaa glyserolitasoon? Tarkastele kysymystä sopivan tilastollisen testin avulla (aseta nollahypoteesi ja tälle vastahypoteesi, sekä määritä testimuuttujan arvo ja sitä vastaava p-arvo). Miten tulkitset testin tuloksen kun kerrot siitä kyselemään tulleelle paikallislehden toimittajalle? (Tenttitehtävä 6.3.2007)

Ratkaisu. Kiinostuksen kohteena on tutkimuksessa havaintojen X_2 ja X_1 erotus $D = X_2 - X_1$ kullekin koehenkilölle. Jotta erotusta voidaan tutkia tilastollisesti, täytyy sille olettaa jokin jakauma. Luonnollinen vaihtoehto on olettaa mittaustuloksille X_1 ja X_2 kaksiulotteinen normaalijakauma, jolloin myös D:llä on normaalijakauma. Merkitään D:n jakauman tuntemattomia odotusarvo- ja varianssiparametreja symboleilla μ ja σ^2.

Kysymystä pyritään selvittämään asettamalla yksinkertainen hypoteesi, jonka mukaan "liikunta ei vaikuta kolesterolitasoon". Tälle vastakkainen väite on siten "liikunta vaikuttaa kolesterolitasoon", jolloin vaikutuksen suunta — ellei siitä ole muulla perusteella selvää ennakkokäsitykstä — voisi periaatteessa olla joko ylös- tai alaspäin. Nollahypoteesiksi asetetaan siten $H : \mu = 0$ ja vaihtoehtoiseksi hypoteesiksi $A : \mu \neq 0$. Koska parametrit μ ja σ^2 ovat tuntemattomia, voidaan hypoteesia testata käyttäen t-testiä. Testisuure määräytyy kaavan (4.6) mukaisesti

$$t_{n-1}(\boldsymbol{X}_n) = \frac{\bar{X}_n - \mu}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}/\sqrt{n}}$$

ja noudattaa $t(n-1)$-jakaumaa. Sijoittamalla tähän kaavaan arvot $n = 11$, $\mu = 0$, $\bar{x}_{11} = \frac{237}{11}$ ja $\frac{1}{10}\sum_{i=1}^{11}(x_i - \bar{x}_{11})^2 \approx 19.08$ saadaan testisuureen arvoksi

$$t_{10}(\boldsymbol{x}_n) \approx \frac{\frac{237}{11}}{(19.08)/\sqrt{11}} \approx 3.745.$$

Taulukon persuteella $t(10)$-jakauman 0.995:n kvantiilipiste on 3.16927 ja 0.9995:n kvantiilipiste 4.5869, joiden väliin testisuureen arvo osuu. Koska kyseessä on kaksisuuntainen testi, sijaitsee testin p-arvo välillä $(0.001, 0.01)$. Näin ollen nollahypoteesi voitaisiin hylätä 0.01:n merkitsevyystasolla.

Huom. Tehtävässä ei voida käyttää kahden odotusarvon vertailuun tarkoitettua testiä, jonka testisuuren määräytyy kaavalla (4.8). Testisuure ei noudattaisi t-jakaumaa, koska muuttujat X_1 ja X_2 eivät ole riippumattomia.

Harjoitustehtävä 4.1 *Näytä, että testattaessa odotusarvoa koskevaa hypoteesia $H : \mu = \mu_0$ yksisuuntaisen testin (vastahypoteesina $A : \mu > \mu_0$) voimakkuus on suurempi kuin saman merkitsevyystason kaksisuuntaisen testin (vastahypoteesina $A : \mu \neq \mu_0$). Tarkastele esimerkkinä tason $\alpha = 0.05$ t-testiä silloin kun otoksesta määritetty keskihajonta $s = 1$ ja otoskoko $n = 30$.*

Harjoitustehtävä 4.2 *Tutkittaessa E.S.P-kyvyn* (extra-sensory perception) *olemassaolon mahdollisuutta erästä koehenkilöä pyydettiin arvaamaan toisessa huoneessa yksi kerrallaan nostettavien korttien väri (sininen tai punainen). Kokeessa nostettiin yhteensä* 50 *korttia hyvin sekoitetusta suuresta pakasta, jolloin koehenkilö veikkasi* 30 *kortin värin oikein. Antaako saatu koetulos "tilastollisesti merkitsevää" tukea hypoteesille E.S.P.-kyvyn olemassaolosta? [Opastus: Käytä* $\mathrm{Bin}(50, 0.5)$*-jakaumalle normaalijakauma-approksimaatiota.]*

Harjoitustehtävä 4.3 *Eräässä amerikkalaisessa yliopistossa tehdyssä tutkimuksessa todettiin, että* 50 *aktiivisesti urheilevan miesopiskelijan pituuden keskiarvo oli* 88.2 *tuumaa ja keskihajonta* 2.5 *tuumaa. Vastaavat luvut* 50 *urheilua harrastamattoman opiskelijan ryhmästä mitattuina olivat* 87.5 *tuumaa ja* 2.8 *tuumaa. Oletetaan, että pituuksien todellinen vaihtelu näiden ryhmien sisällä noudattaa likimäärin normaalijakaumaa ja että varianssit ryhmien sisällä ovat yhtä suuria. Testaa tämän oletuksen pätiessä hypoteesia, että ryhmien välillä ei ole systemaattista pituuseroa ja määritä testin tuloksena saatava p-arvo. Minkä johtopäätöksen teet tällä perusteella?*

Harjoitustehtävä 4.4 *(Jatkoa edelliseen tehtävään.) Kuinka suuri otos kumpaankin ryhmään tulisi kerätä, jotta havaittu* 0.7 *tuuman ero keskiarvoissa olisi tilastollisesti merkitsevä merkitsevyystasolla* 0.05*?*

Harjoitustehtävä 4.5 *Tutkittaessa eräässä kokeessa oppimistuloksiin vaikuttavia tekijöitä yhteensä* 12 *koehenkilöä sijoitettiin arpomalla kahteen eri valmennusryhmään. Järjestetyssä kokeessa heille saatiin seuraavat pistemäärät:*

- *Menetelmä 1: 20, 17, 10, 25, 24, 22, 15;*
- *Menetelmä 2: 26, 31, 23, 35, 20.*

Tukeeko saatu tulos väitettä, että valmennusmenetelmien tehokkuudessa on eroa? Määritä testin tuloksena saatava p-arvo kaksisuuntaisen vaihtoehtoisen hypoteesin tapauksessa olettaen, että pistemäärät jakautuvat likimäärin normaalisesti ja että ja että niiden vaihtelu koeryhmien sisällä on samaa suuruusluokkaa.

Liite A

Vanhoja tenttitehtäviä

Tehtävä A.1 *Eräässä perheiden käytettävissä olevan tulon määrää koskevassa tutkimuksessa tiedusteltiin erikseen kummankin työssä käyvän puolison ansiotuloja. (Tutkimus koski vain perheitä, joissa molemmat puolisot kävivät ansiotyössä.) Tulkitaan miehen ansiotulon määrä satunnaismuuttujaksi X ja vastaavasti vaimon ansiotulon määrä satunnaismuuttujaksi Y. Oletetaan, että kyselyn perusteella on saatu selville tarkat arvot näiden muuttujien sekä odotusarvoille että variansseille. (a) Onko perusteltua ajatella, että puolisoiden yhteenlasketun ansiotulon odotusarvo on muuttujien X ja Y odotusarvojen summa? Perustele vastauksesi. (b) Onko perusteltua ajatella, että puolisoiden yhteenlasketun ansiotulon määrän varianssi on muuttujien X ja Y varianssien summa? (Tenttitehtävä 9.8.2006)*

Tehtävä A.2 *Neljän opiskelijapojan painot ovat 70 kg, 73 kg, 81 kg ja 88 kg. Heistä arvotaan satunnaisesti joku ja merkitään tämän painoa kirjaimella X. Määritä X:n jakauma (pistetodennäköisyydet), odotusarvo ja varianssi. (Tenttitehtävä 9.5.2007)*

Tehtävä A.3 *Olet kiinnostunut arvioimaan todennäköisyyksiä, joilla lasipurkkiin laittamasi nasta laskeutuu helistyksen jälkeen "selälleen" tai vastaavasti "kyljelleen". Päätät suorittaa tätä varten neljä koesarjaa, joissa kussakin lasket tarvittavien helistysten lukumäärän kunnes siinä saadaan ensimmäisen kerran tulos "kyljelleen". Kuvaat kussakin koesarjassa tarvittavien helistysten lukumäärää geometrisen jakauman avulla, jonka pistetodennäköisyydet määritellään $P(X = k) = (1 - \theta)^{k-1}\theta$, $k = 1, 2, \ldots$, jolloin parametri θ voidaan tulkita "kyljelleen laskeutumisen todennäköisyydeksi". Suorittamissasi koesarjoissa tarvitset ensimmäisessä yhteensä kolme, toisessa yhden, kolmannessa kuusi ja neljännessä viisi helistystä. Johda näitä havaintoja vastaava uskottavuusfunktion lauseke sekä sen perusteella θ:n suurimman uskottavuuden estimaatti. (Tenttitehtävä 17.5.2005)*

Tehtävä A.4 *Olkoon $X_1, X_2, \ldots, X_n$ riippumaton otos $\mathrm{Exp}(\lambda)$-jakautuneita satunnaismuuttujia mallissa $\boldsymbol{M}$, tiheysfunktiona $f_X(x; \lambda) = \lambda \exp(-\lambda x)$, $x > 0$. Johda parametrin λ suurimman uskottavuuden estimaattorin lauseke. (Tenttitehtävä 15.6.2005)*

Tehtävä A.5 *Junan aikataulun mukainen saapumisaika määräasemalle on klo 12.00. Käytännössä on kuitenkin havaittu, että sen todellinen saapumisaika noudattaa likimäärin normaalijakaumaa, jonka odotusarvo on klo 12.04 ja keskihajonta 3 minuuttia. Määritä tällä perusteella todennäköisyys, että (a) juna myöhästyy aikataulusta, mutta korkeintaan yhden minuutin, (b) juna myöhästyy enemmän kuin 5 minuuttia. (Tenttitehtävä 10.8.2005)*

Tehtävä A.6 *Olkoot $X_1, X_2, \ldots, X_n$ riippumattomia muuttujia binomikokeessa, jonka parametria $\theta = P(X_i = 1)$ halutaan estimoida saatujen havaintojen $x_1, x_2, \ldots, x_n$ perusteella ja käyttäen Bayes-päättelyä. Näytä, että jos prioriksi valitaan Beta(α, β)-jakauma (tiheysfunktiona $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, myös posteriorijakauma on Beta-jakauma. Mitkä sen parametrit ovat? (Tenttitehtävä 14.6.2006)*

Tehtävä A.7 *Liisa on hankkinut itselleen kanan ja hän on kiinnostunut tietämään, minkä kokoisia munia se munii. Liisa on opiskellut hieman tilastotiedettä ja hän arvelee, että munien painoa voitaisiin riittävän hyvin kuvata yksinkertaisella otannalla normaalijakaumasta $N(\mu, \sigma^2)$, missä parametrit μ ja σ^2 ovat kuitenkin tuntemattomia. Niinpä Liisa päättää punnita seuraavien 10 munan painon, jolloin hän saa tulokseksi seuraavat lukemat: 56, 62, 55, 59, 62, 60, 61, 57, 58 ja 60 grammaa. Liisa päättää määrittää aineiston perusteella tason 0.95 luottamusvälin painojen odotusarvolle μ. Mitä hän saa tulokseksi? (Tenttitehtävä 15.6.2005)*

Tehtävä A.8 *Kesätyöpaikassasi joudut kuvaamaan esimiehellesi eräässä tutkimusraportissa saatuja tuloksia. Raportissa kerrotaan, että kahden tutkitun menetelmän välille "saatiin tilastollisesti merkitsevä ero tasolla $\alpha = 0.05$". Esimiehesi kysyy sinulta, mitä tämä oikeastaan tarkoittaa. Neuvonpitoon osallistuva työkaverisi vastaa silloin: "Se tarkoittaa sitä, että nollahypoteesi H, jonka mukaan menetelmien välillä ei ole eroa, on tosi vain todennäköisyydellä 0.05." Hyväksytkö hänen vastauksensa? Ellet hyväksy, esitä perustelu tälle mielipiteellesi. Miten olisit itse vastannut? (Tenttitehtävä 12.11.08)*

Tehtävä A.9 *Eräällä sahalla mitattiin kymmenen satunnaisesti valitun lankun paksuuksiksi 2.11, 2.02, 2.20, 1.98, 2.14, 1.99, 2.09, 2.19, 2.23 ja 2.10 tuumaa. Olettaen, että paksuudet noudattavat likimäärin normaalijakaumaa, määritä tällä perusteella sen odotusarvolle tason $1-\alpha = 0.95$ luottamusväli. Voitko suoraan määrittämäsi luottamusvälin perusteella päätellä, tulisiko odotusarvoa koskeva hypoteesi $H_0 : \mu = 2.00$ hyväksytyksi tai hylätyksi kaksipuolisessa tason $\alpha = 0.05$ testissä? Perustele väitteesi. (Tenttitehtävä 10.8.2005)*

Tehtävä A.10 *Eräs savukefilttereiden valmistaja väittää, että heidän filtterinsä suodattavat tehokkaammin savukkeissa olevaa nikotiinia kuin yleisesti käytössä oleva vertailuvalmiste. Koska eri savukemerkit eroavat jonkin verran toisistaan, näitä kahta filtterityyppiä verrattiin määrittäen yhteensä yhdeksän savukemerkin tapauksessa filtteriin kertyneiden nikotiinimäärien erotus uuden tyyppisen ja vertailuvalmisteen välillä. Erotusten keskiarvo oli $\mu = 1.32$ mg ja keskihajonta $s = 2.35$ mg. Muotoile väitettä vastaava testiongelma (aseta hypoteesi H ja tämän vastahypoteesi A) ja arvioi testin tulosta käyttäen siitä saatavaa p-arvoa. (Tenttitehtävä 9.5.2006)*

Vihjeitä tehtäviin

A.3 Estimaattorin lausekkeen määrittämiseksi tulee etsiä parametrin arvo, joka maksimoi uskottavuusfunktion. Tämä on helpointa tehdä tutkimalla uskottavuusfunktion logaritmia ja sen derivaattaa.

A.4 Kuten A.3.

A.5 Käytä kaavaa (2.11) todennäköisyyksien laskemiseen.

A.6 Muodosta parametrin θ posteriorijakauma käyttäen kaavaa (2.21) ja kerää samanlaiset termit yhteen. Huomaa, että voit jättää pois termit, jotka eivät sisällä θ:aa.

A.7 Laske estimaatit odotusarvolle μ ja σ^2 ja käytä kaavaa (3.9).

A.9 Vertaa luottamusväliä (3.9) kaavan (4.7) kaltaiseen nollahypoteesin hyväksymisalueeseen.

A.10 Käytä yksisuuntaista t-testiä.