2.1 Havaintoja vastaava todennäköisyysmalli

- Numeerinen aineisto (engl. *data*) $y_1, \ldots, y_n$, jossa kukin y_i on jokin tunnettu luku.
- Havaintojen lukumäärä n on nimeltään otoskoko (engl. *sample size*).
- Ennen havaintojen tekoa aineiston arvot ovat epävarmoja (mittausvirheiden, koetilanteessa tehdyn satunnaistamisen, populaation luonnollisen vaihtelun tms. syyn takia). Kokeen tai otannan toistaminen voisi tuottaa toisenlaiset havainnot.

Perusajatus

- Mallinnamme: arvot $y_1, \ldots, y_n$ ovat satunnaismuuttujien $Y_1, \ldots, Y_n$ toteutuneita arvoja (eli niiden reaalisaatioita).
- Kukin satunnaismuuttuja Y_i on kuvaus $Y_i : \Omega \to \mathbb{R}$.
- Reaalisaatio tarkoittaa sitä, että

$$y_1 = Y_1(\omega^{\text{act}}), y_2 = Y_2(\omega^{\text{act}}), \ldots, y_n = Y_n(\omega^{\text{act}}), \qquad (1)$$

  jossa $\omega^{\text{act}} \in \Omega$ on todennäköisyysmallissa aktualisoitunut alkeistapaus.
- Vektorimerkinnät:

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

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

Parametrinen malli

- Oletamme, että kaikki satunnaismuuttujat Y_i ovat joko diskreettejä (pistetodennäköisyysfunktio, ptnf) tai että niillä kaikilla on jatkuva jakauma (tiheysfunktio, tf).
- Parametrisessa mallissa oletamme, että yhteisjakauman yptnf/ytf (y = yhteis-) tunnetaan muuten, mutta jokin sen parametri (tai useampi parametreja) on tuntematon:

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

- Yllä $\mathbf{y}$ on vapaa muuttuja (ei tässä vaiheessa vielä aineisto).

Yptnf/ytf — riippumattomat satunnaismuuttujat

Jos satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat riippumattomia, kun parametrin arvo on kiinnitetty, niin yptnf/ytf on tulomuotoa

$$\begin{aligned} f(\mathbf{y}; \theta) &= f_{Y_1}(y_1; \theta)\, f_{Y_2}(y_2; \theta) \cdots f_{Y_n}(y_n; \theta) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i; \theta) \end{aligned} \tag{2}$$

Tässä $f_{Y_i}(u; \theta)$ on satunnaismuuttujan Y_i ptnf/tf, kun parametrinarvo on θ.

Yptnf/ytf — satunnaisotos

Jos satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat riippumattomia ja lisäksi samoin jakautuneita, ja niiden jakaumalla on ptfn/tf $g(y; \theta)$, niin niiden yptnf/ytf on

$$f(\mathbf{y}; \theta) = g(y_1; \theta) \cdots g(y_n; \theta) = \prod_{i=1}^{n} g(y_i; \theta). \tag{3}$$

Tällöin sanotaan, että satunnaismuuttujat

$$Y_1, \ldots, Y_n$$

ovat satunnaisotos (engl. *random sample*) ko. jakaumasta.

Tilastollinen päättely parametrisessa mallissa

- Satunnaisvektorin $\mathbf{Y}$ jakauma tunnetaan, jos parametrinarvo tunnetaan.
- Parametrinarvo on tuntematon, joten sitä yritetään arvioida eli estimoida.
- Parametrista tiedetään ainakin sen verran, että osataan sanoa, missä joukossa Θ (parametriavaruus) sen arvot voivat olla.

2.2 Pallot kulhossa

- Kulhossa on samankokoisia ja samasta materiaalista valmistettuja valkoisia ja mustia palloja yhteensä N (tunnettu lukumäärä).
- Valkoisten pallojen lukumäärä on θ (tuntematon parametri),

$$\theta \in \{0, 1, \ldots, N\}.$$

- Mustien pallojen lukumäärä on $N - \theta$.
- Kulhoa ravistetaan tarmokkaasti, ja sitten siitä nostetaan yksi pallo sokkona.
- Luonnollinen malli:

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

Havaintoja vastaavat satunnaismuuttujat

- Kulhosta nostetaan satunnaisesti pallo n kertaa, nostettu pallo palautetaan aina kulhoon. Kulhoa ravistetaan aina perusteellisesti ennen nostoa.

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

- Yhden muuttujan Y_i ptnf on

$$P_\theta(Y_i = 1) = \theta/N$$
$$P_\theta(Y_i = 0) = 1 - \theta/N.$$

- Sama yhdellä kaavalla:

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

Yhteispistetodennäköisyysfunktio

- Koska kulhoa aina ravistetaan perusteellisesti ennen kutakin nostoa ja koska nostetut pallot aina palautetaan kulhoon, niin on luonnollista ajatella, että nostoja vastaavat satunnaismuuttujat ovat riippumattomia.
- Arkijärjen mukaan tieto yhden noston lopputuloksesta ei voi vaikuttaa toisen noston todennäköisyysjakaumaan.
- Siis yptnf on tulomuotoa (1). Jokaisella Y_i on sama jakauma, joten kyseessä on satunnaisotos (2).

Yptnf (pallot kulhossa)

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

Kukin y_i saa joko arvon 0 tai 1 ja parametri θ on jokin luvuista $0, 1, \ldots, N$. Tästä:

$$
f(\mathbf{y};\theta) = \left(\frac{\theta}{N}\right)^{t(\mathbf{y})} \left(1-\frac{\theta}{N}\right)^{n-t(\mathbf{y})}, \tag{4}
$$

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

Toinen parametrointi

- Malli voidaan aina parametroida useilla eri tavoilla.
- Valitaan parametriksi ϕ valkoisten pallojen suhteellinen osuus kulhossa olevista palloista, eli

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

- Aineistoa vastaavan satunnaisvektorin $\mathbf{Y}$ yptnf on nyt

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

- Uutta parametrointia vastaava parametriavaruus on joukko

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

Parametrin olemus (pallot kulhossa)

- Tässä esimerkissä parametrilla on konkreettinen tulkinta. Tämä on harvinaista tilastollisissa malleissa.
- Parametrin todellinen arvo voitaisiin selvittää katsomalla kulhoon.
- Kokeen lopputuloksen perusteella saattaa olla mahdollista sulkea pois tiettyjä parametrinarvoja. Jos yhdessäkin nostossa saadaan valkoinen pallo, niin arvo $\theta = 0$ voidaan sulkea pois. Jos yhdessäkin nostossa saadaan musta pallo, niin arvo $\theta = N$ voidaan sulkea pois.
- Kuvatun kokeen puitteissa parametrin todellista arvoa ei kuitenkaan voida selvittää täysin varmasti (mikäli $N \geq 3$).

Miksi päädyttiin riippumattomuuteen?

- Kulhoa ravistettiin perusteellisesti ennen pallon nostoa sokkona.
- Nostettu pallo palautettiin aina kulhoon.
- Jos nämä ehdot eivät ole voimassa, niin edellä käsitelty malli ei ole hyvä kuvaus todellisuudelle.
- Monisteessa käsitellään hieman tilannetta, jossa palloja ei palauteta kulhoon noston jälkeen.

2.3 Nasta purkissa

- Purkissa on nasta. Purkkia ravistetaan tarmokkaasti, ja sitten merkitään muistiin, laskeutuuko nasta selälleen vai kyljelleen. Tätä koetta toistetaan n kertaa.
- Otamme käyttöön satunnaismuuttujat Y_i siten, että

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

Nasta purkissa — parametrointi

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

Nasta purkissa — yptnf

- Ajattelemme, että eri ravistusten jälkeiset lopputulokset ovat keskenään riippumattomia, koska arkijärjen mukaan tieto yhden ravistuksen lopputuloksesta ei voi vaikuttaa toisen ravistuksen lopputuloksen todennäköisyysjakaumaan.
- Päädymme yhteispistetodennäköisyysfunktioon

$$f(\mathbf{y};\theta) = \theta^{y_1}(1-\theta)^{1-y_1}\cdots\theta^{y_n}(1-\theta)^{1-y_n} = \theta^{t(\mathbf{y})}(1-\theta)^{n-t(\mathbf{y})}, \tag{5}$$

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

2.4 Binomikoe

Molemmat esimerkit ovat erikoistapauksia ns. binomikokeesta:

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

Yptnf binomikokeessa

Binomikokeessa satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteisjakaumalla on yptf

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

jossa

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

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

Onnistumisten lukumäärän jakauma

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

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

kertomatta, missä järjestyksessä onnistumiset ja epäonnistumiset sattuivat.

- Jos onnistumisten lukumäärää pidetään satunnaismuuttujana ts. jos käsitellään satunnaismuuttujaa

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

niin tällöin X noudattaa tunnetusti binomijakaumaa parametreilla n ja p.

Binomijakauma

- Lyhyemmin merkittynä

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

jossa n on toistojen lukumäärä (tai otoskoko), ja $0 \leq p \leq 1$ on onnistumistodennäköisyys (ykkösen todennäköisyys) yhdessä kokeessa.

- Binomijakauman pistetodennäköisyysfunktio on

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \quad (6)$$

2.5 Kaksi lähestymistapaa

- Parametrisessa mallissa havaintoja vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma tunnetaan täysin, jos mallin $f(\mathbf{y}; \theta)$ parametrin θ arvo tunnetaan.
- Tilastollisessa päättelyssä θ on tuntematon luku.
- Pyrkimyksenä on arvioida eli estimoida parametrin θ arvoa havaitun aineiston $\mathbf{y}$ perusteella, ja yrittää vielä kuvailla tähän arvioon liittyvää epävarmuutta.
- Bayesiläinen päättely (engl. *Bayesian inference*) vs. frekventistinen päättely (engl. *frequentist inference*).

Bayesiläinen päättely — vähän historia

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

Frekventistinen päättely — vähän historiaa

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

Nykytilanne

- Bayesiläinen lähestymistapa alkoi tulla uudestaan suosituksi vasta 1980-luvun loppupuolelta lähtien.
- Uusi nousu perustui suurelta osin uusiin laskentamenetelmiin sekä siihen, että tietokoneiden käyttö alkoi niihin aikoihin tulla jokapäiväiseksi.
- Nykyään useimmat (ammatti-)tilastotieteilijät vähintäänkin ymmärtävät, mistä kumassakin lähestymistavassa on kyse.
- Soveltajille opetetaan tyypillisesti yksinomaan frekventististä päättelyä — bayesiläinen päättely vaatii vähän laajempia tietoja todennäköisyyslaskennasta kuin mitä frekventistinen.

2.5.1 Frekventistinen lähestymistapa

- Frekventistisessä lähestymistavassa parametri θ on tuntematon, mutta kiinteä (eli ei-satunnainen) luku.
- Parametrista tiedetään ainoastaan se, missä joukossa eli parametriavaruudessa sen arvot voivat olla.
- *Tilastollinen malli* koostuu satunnaisvektorin $\mathbf{Y}$ jakauman yptnf:stä tai ytf:stä $f(\mathbf{y};\theta)$ sekä parametriavaruudesta Θ.
- Tilastollinen malli on (frekventistille) jakaumien

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

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

Tn-laskennan rooli frekventistille

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

Frekventistisen päättelyn erityiskysymyksiä

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

Väliestimointi. Parametriavaruudesta rajataan sellainen väli (tai joukko), jonka luotetaan sisältävän oikean parametrin arvon. Tällä tavalla kuvataan piste-estimoinnissa saatavaa tarkkuutta.

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

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

2.5.2 Bayesiläinen lähestymistapa

- Nyt myös parametri tulkitaan satunnaismuuttujaksi.
- Aineistoa vastaavan satunnaisvektorin jakauma $f(\mathbf{y}; \theta)$ ymmärretään satunnaisvektorin $\mathbf{Y}$ ehdolliseksi jakaumaksi, kun parametrilla on arvo θ.
- Käytetään ehdollisen jakauman merkintää $f(\mathbf{y} \mid \theta)$.
- Kaikki koetilanteeseen liittyvä taustatieto pyritään esittämään parametrin priorijakaumana, joka on todennäköisyysjakauma parametriavaruudessa.
- Priorijakauma esittää kvantitatiivisesti tutkijan epävarmuuden parametrin oikeasta arvosta ennen (lat. *a priori*) kuin havaintoa on tehty.

Tilastollinen malli bayesiläiselle

- Bayesiläisessä lähestymistavassa *tilastollinen malli* koostuu ehdollisesta jakaumasta $f(\mathbf{y} \mid \theta)$ sekä priorijakaumasta.
- Priorijakauma ja havaintovektorin $\mathbf{Y}$ ehdollinen jakauma määräävät näiden kahden satunnaissuureen yhteisjakauman todennäköisyyslaskennan sääntöjen mukaan.

Bayesiläinen päättely pähkinänkuoressa

- Parametrin ja havaintovektorin $\mathbf{Y}$ yhteisjakaumasta siirrytään parametrin posteriorijakaumaan eli parametrin ehdolliseen jakaumaan, kun tiedetään, että $\mathbf{Y}$ on saanut arvon $\mathbf{y}$.
- Posteriorijakauma määräytyy periaatteessa automaattisesti todennäköisyyslaskennan sääntöjen avulla, mutta käytännössä sen ominaisuuksia joudutaan usein selvittämään raskaiden laskujen avulla.
- Posteriorijakauma esittää kvantitatiivisesti tutkijan epävarmuuden parametrin arvosta, kun havainto otetaan huomioon.
- Usein myös bayesiläisessä päättelyssä lasketaan piste-estimaatteja ja väliestimaatteja, vaikka ne ovatkin vain eräitä (varsin köyhiä) tapoja kuvailla posteriorijakaumaa.

2.5.3 Yhteenveto

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