Luku 6 - Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma

Tiivistelmä

Luvun ydinviesti

Tässä luvussa tarkastellaan otoksia ja otosjakaumia "tilastollisemmin" mitä edellisten erityisesti otantaa koskevan johdannan yhteydessä, toisin sanoen nyt tutustumme kevyesti tilastolliseen päättelyyn!

Tämä luku on kurssin materiaalista matemaattisin, mutta siitä ei tule huolestua. Tavoite on yhdistää aiemmin opittuja konsepteja matemaattiseen merkintätapaan.

Tilastolliseen päättelyyn perehdytään tarkemmin tilastollisen päättelyn peruskurssilla (TILM3555)

Ydinviestinä olkoon siis matemaattisen formaalin esitystavan yhdistäminen aiemmin opittuihin konsepteihin! Uutena asiana puhutaan hieman aineistosta laskettaviin tunnuslukuihin ja estimaattoreihin.

Satunnaisotos, yhteisjakauma ja tilastollinen malli

Satunnaismuuttujilla on todennäköisyysjakaumat, joita tilastotieteessä kuvataan todennäköisyys- eli tiheysfunktion (tai pistetodennäköisyysfunktion) avulla.

Satunnaisotos
Olkoot $Y_1, \ldots, Y_n$ riippumattomia ja samoinjakautuneita satunnaismuuttujia, joiden tiheysfunktiota (tf., tai pistetodennäköisyysfunktiota (ptnf)) merkitään $f(y, \theta)$:llä, jossa y on yksittäisen sm:jan Y:n realisaatio ja θ on jokin jakauman muodon määräävä parametri (tai parametrit).
Parametrin θ arvoa ei yleensä tunneta ja tavoitteena onkin päätellä, **estimoida**, sen arvo lopulta käytettävissä olevasta aineistosta.

Satunnaisotoksen tilastollinen malli

Satunnaismuuttujien havaitut arvot muodostavat siis satunnaisotoksen. Ne ovat kiinteitä lukuja, mutta vaihtelevat satunnaisesti otoksesta toiseen.

Täten satunnaisotannassa **satunnaisuus liittyy siis havaintoarvojen vaihteluun satunnaisesti otoksesta toiseen.** **Tilastollinen malli** on näiden havaintoarvojen otosten välistä vaihtelua kuvaava **yhteisjakauma**.

Edellä oletettiin että satunnaismuuttujat ovat keskenään riippumattomia, jolloin tämä yhteisjakauma on seuraavaa tulomuotoa $f(y_1, \ldots, y_n; \theta) = f(y_1; \theta) \times \cdots \times f(y_n; \theta)$

Tilastollinen mallin muoto riippuu tutkijan tekemästä jakaumaoletuksesta, jonka monimutkaisuus ilmenee sen parametrien määrästä. Niin sanotun **parsimoonisuus-** eli **vähäparametrisuusperiatteen** mukaan tilastollisen mallin tulisi olla niin yksinkertainen, eli vähäparametrinen, kuin mahdollista satunnaisotoksen riittävään kuvaamiseen.

Otosjakauma: klassisen tilastotieteen näkökulma

Ajatus aineiston keruun toistamisesta ilmentää klassisen (frekventistisen) tilastotieteen käsitystä tilastollisesta stabiliteetista, jolle tilastollinen päättely perustuu.

Keskeistä tälle on tilastolliseen malliin $f(y_1, \ldots, y_n; \theta)$ liitetty jakaumaoletus. Mikäli kerättäisiin uusi satunnaisotos, eli toistettaisiin aineiston keruu, niin satunnaismuuttujien havaitut arvot, havaintoarvot, vaihtelisivat tilastollisen mallin jakauman kuvaamalla tavalla!

Otosjakauma: Estimaattori ja estimaatti

Satunnaisotoksesta voidaan laskea erilaisia **tunnuslukuja/otossuureita,** joita merkitään $T = g(Y_1, \ldots, Y_n)$, ts. ne ovat aineiston funktioita. Tunnusluvut ovat **satunnaismuuttujien funktioina myös satunnaismuuttujia.**

Tunnusluvuilla on täten myös havaittu arvo satunnaisotoksen määräämässä pisteessä, merkitään $t = g(y_1, \ldots, y_n)$. Tilastollisen tutkimuksen tavoite on pyrkiä aineiston avulla arvioimaan, estimoimaan, tunnusluvun nk. **todellinen arvo**, $g(\theta)$.

Koska tunnusluku/estimaattori T on satunnaismuuttuja, sillä on todennäköisyysjakauma, jota kutsutaan tunnusluvun T otosjakaumaksi ja joka myös riippuu tuntemattomista parametreista.

Keskeiset termit: harhattomuus

Harhattomuus
Estimaattorin odotettavissa oleva arvo yhtyy tuntemattoman parametrin todelliseen arvoon eli $E(\hat{\theta}) = \theta$.

- Harhaton estimaattori tuottaa keskimäärin oikean kokoisia arvoja (estimaatteja) estimoitavalle parametrille.
- Estimaattorin tuottama arvo, estimaatti, parametrille saattaa vaihdella otoksesta toiseen paljonkin, mutta odotusarvon frekvenssitulkinnan mukaan otoskohtaiset estimaatit jakautuvat otantaa toistettaessa (symmetrisesti) parametrin todellisen arvon ympärille.

Keskeiset termit: tyhjentävyys, tehokkuus ja tarkentuvuus

Tyhjentävyys
Tyhjentävä estimaattori käyttää kaiken otokseen sisältyvän parametria θ koskevan informaation.

Tehokkuus
Kahdesta saman parametrin θ estimaattorista tehokkaampi on se, jonka varianssi on pienempi.

Tarkentuvuus
Tarkentuvan estimaattorin $\hat{\theta}$ arvot lähestyvät parametrin θ oikeaa arvoa otoskoon kasvaessa.

Otoskeskiarvo estimaattorina

Olkoon $Y_1, \ldots, Y_n$ riippumattomia ja samoinjakautuneita sm:jia, ja että kyseisen jakauman odotusarvo on $E(Y_i) = \mu$ ja varianssi $\text{Var}(Y_i) = \sigma^2$.

- Satunnaismuuttujien (aritmeettinen) keskiarvo on $\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ ja sitä vastaa otoskeskiarvo $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Kun satunnaismuuttujat ovat samoin jakautuneet odotusarvonaan μ, on otoskeskiarvo jakauman odotusarvon harhaton estimaattori, ts. $E(\bar{Y}) = \mu$.

Tällöin keskiarvo $\bar{Y}$ on tunnusluku, jota käytetään odotusarvon estimoimiseen, eli se on estimaattori ja täten sillä on myös jakauma, jota kuvaa odotusarvo $E(\bar{Y}) = \mu$ ja varianssi $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$. Huomataan, että otoskoon kasvaessa keskiarvon otosjakauma keskittyy yhä voimakkaammin odotusarvon ympärille.

Otosvarianssi estimaattorina

Perusjoukon tasolla sm:jien vaihtelua kuvataan **populaatiovarianssilla** $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu)^2$ ja sen aineistosta laskettava vastine on **otosvarianssilla** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Huomaa että **otosvarianssi** on eri asia kuin **otoskeskiarvon varianssi.**

Otoskeskiarvo ja otosvarianssi ovat siis satunnaismuuttujia, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen. Näitä tunnuslukuja käytetään arvioimaan perusjoukon, eli populaation, odotusarvoa ja varianssia, joten ne ovat myös estimaattoreita.

Normaalijakautunut otos

Jos satunnaisotoksen muodostavat havainnot $Y_1, \ldots, Y_n$ ovat peräisin normaalijakaumasta, niin voidaan osoittaa että havaintojen keskiarvo on myös normaalisti jakautunut, merkitään $\bar{Y} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.

Ns. **asymptoottiseen teoriaan** vedoten voidaan osoittaa että tämä pätee myös ilman normaalisuusoletusta kun otoskoko on suuri.

Esimerkki

Standardoitu keskiarvo saadaan vähentämällä keskiarvosta odotusarvo ja jakamalla se keskipoikkeamalla, merkitään $Z = \frac{\bar{Y} - E(\bar{Y})}{D(\bar{Y}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$. Tällöin Z:n odotusarvo $E(Z) = 0$ ja varianssi $\mathrm{Var}(Z) = 1$. Lisäksi, jos $Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, niin Z noudattaa standardoitua normaalijakauma $Z \sim N(0, 1)$.

Suhteellisen frekvenssin otosjakauma

Ohitetaan tiivistelmässä.

Muita tunnuslukuja

Tilastollisessa tutkimuksessa hyödynnetään myös paljon muita otoksesta laskettavia tunnuslukuja. Erilaiset tunnusluvut kertovat muuttujan jakaumasta erilaisia asioita, joiden avulla taustalla olevan satunnaisilmiön luonnetta on helpompi hahmottaa.

- **Moodi** eli tyyppiarvo on havaintoaineiston yleisin muuttujan arvo tai se luokka, jolla on suurin frekvenssi.
- **Mediaani** on järjestetyn havaintoaineiston keskimmäinen arvo. Puolet arvoista on mediaania pienempiä ja puolet suurempia.
- **Fraktiili** on vastaavasti jotain prosenttiosuutta vastaava arvo, joka jakaa aineiston kahteen osaan siten että kyseistä fraktiilia vastaavaa havaintoarvoa pienempiä arvoja on kyseinen prosenttiosuus.
  - Esimerkkeinä **kvartiilit,** jotka jakavat aineiston 25% välein, ts. Q_1 on ns. alakvartiili, joka vastaa 25% fraktiilia ja yläkvartiili Q_3 on 75% fraktiili. Näiden erotusta kutsutaan **kvartiiliväliksi** $= (Q_1, Q_3)$.
- **Vinous** ja **huipukkuus** kuvaavat jakauman muotoa, erityisesti sen poikkeamaa normaalijakaumasta.

Luottamusvälit

Satunnaisotoksesta laskettujen tunnuslukujen luotettavuus on tilastollisen mallin parametrien estimoinnissa keskeinen kysymys.

- Otantaan liittyvän satunnaisvaihtelun vuoksi emme voi varmuudella tietää onko otoksesta laskettu parametriestimaatti "lähellä" vai "kaukana" sen todelisesta arvosta.
- Tätä luotettavuutta voidaan arvioida **luottamusvälin** avulla.

Luottamusväli
Luottamusväli on otoksen perusteella määrätty väli, joka tutkijan valitsemalla todennäköisyydellä (luottamustasolla) peittää tarkasteltavan tilastollisen mallin $f(y;\theta)$ parametrin θ tuntemattoman todellisen arvon. Se perustetaan otostunnusluvun, estimaattorin, otosjakaumaan.

Luottamusväli

Luottamustasoa merkitään usein $1-\alpha$:lla, jossa α on **merkitsevyystaso** (**riskitaso**), usein esimerkiksi $\alpha = 0.05$. Luottamustaso tulkitaan niin, että jos **otantaa** jakaumasta $f(y;\theta)$ toistetaan, niin keskimäärin $100 \times (1-\alpha)\%$ otoksista muodostetuista (konstruloiduista) luottamusväleistä peittää parametrin θ todellisen arvon.

Luottamusväli on tunnetumpi kansankieliseltä nimitykseltään **virhemarginaali,** joka on itse asiassa luottamusvälin puolikas. Todellinen parametriarvo kuuluu saadun estimaatin ja virhemarginaalien sisään jäävälle osuudelle.

- Virhemarginaalin suuruuteen vaikuttavat otosasetelma, otoskoko, luottamustaso ja tutkittavan tilastollisen tunnusluvun jakauma.

Normaalijakauman odotusarvon luottamusväli

Tarkastellaan seuraavaksi lyhyesti normaalijakauman odotusarvon luottamusvälejä silloin kun taustalla oleva populaatio on “iso” (ääretön).

Tarkastellaan satunnaisotosta normaalijakaumasta $Y_1, \ldots, Y_n$, missä satunnaismuuttujat ovat riippumattomia ja niille pätee $Y_i \sim N(\mu, \sigma^2)$.

Muistetaan että keskiarvo on normaalijakauman odotusarvoparametrin **harhaton estimaattori** $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Valitaan **luottamustasoksi** $1 - \alpha$, eli α määrää todennäköisyyden, jolla luottamusväli peittää odotusarvon μ todellisen arvon. Yleinen valinta ihmistieteissä on $\alpha = 0.05$ tai $\alpha = 0.1$ vastaten 95% ja 90% luottamustasoa. Luonnotieteissä α on usein paljon pienempi.

Normaalijakauman odotusarvon luottamusväli

Seuraavaksi määrätään **luottamuskertoimet** $-z_{\alpha/2}$ ja $z_{\alpha/2}$, joille pätee $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$, jossa Z on standardoitu satunnaismuuttuja ja noudattaa $N(0,1)$ jakaumaa (standardinormaalijakaumaa).

Sijoitetaan standardoidun satunnaismuuttujan määritelmä yo. epäyhtälöketjuun ja saadaan

$$-z_{\alpha/2} \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

joka voidaan kirjoittaa uudelleen muodossa

$$\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Normaalijakauman odotusarvon $(1-\alpha) \times 100\%$ luottamusväli on siis

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

Luottamustason tulkinta

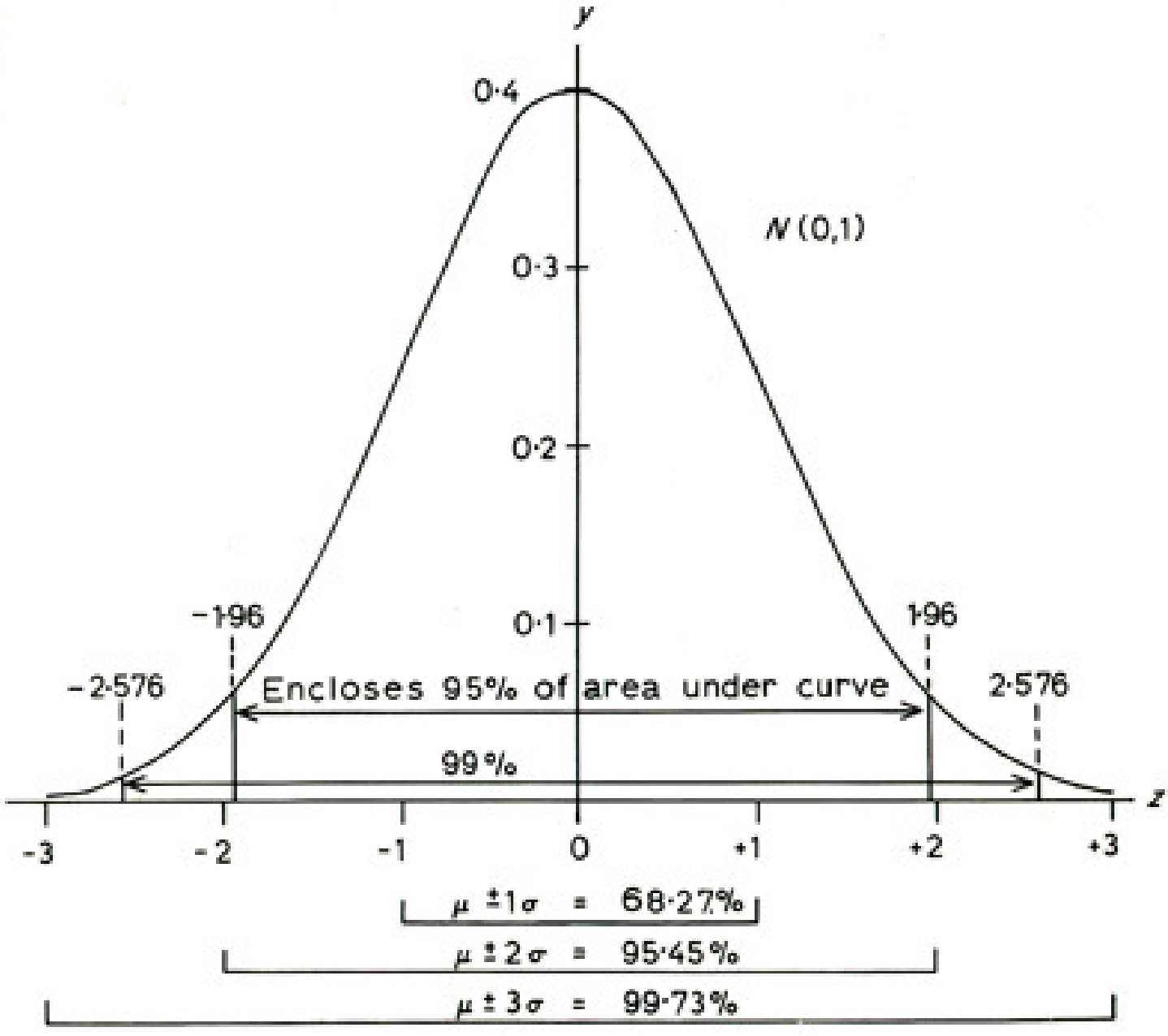
Miten luottamustasoa $(1-\alpha)$ tulisi tulkita? Olisi toivottavaa, että estimoidulle parametrille pystyttäisiin konstruoida mahdollisimman lyhyt luottamusväli, sillä tämä tarkoittaisi että saatu estimaatti olisi luotettavampi!

Samanaikaisesti olisi toivottavaa, että luottamustaso olisi mahdollisimman korkea, sillä tämä tarkoittaa nimensä mukaisesti saatu estimaatti olisi luotettavampi!

Molempien vaatimusten samanaikainen täyttäminen ei ole kuitenkaan mahdollista, jos otoskoko n pidetään kiinteänä:

- Luottamustason kasvattaminen pidentää luottamusväliä, jolloin tieto parametrin μ todellisesta arvosta tulee epätarkemmaksi.
- Luottamusvälin lyhentäminen pienentää luottamustasoa, jolloin tieto parametrin μ todellisesta arvosta tulee epävarmemmaksi.

Luottamustason tulkinta graafisesti



Otoskoko

Kuten jo tiedetään, yksi tilastollisen tutkimuksen (päättelyn) keskeisiä tavoitteita on yleistää otoksen pohjalta tehty päättely koskemaan koko perusjoukkoa.

- Kun otoskoko on liian pieni, voi otos **sattumalta** poiketa paljonkin perusjoukosta.
- Todella suuren otoksen kerääminen/koostaminen voi olla **työlästä, kallista** tai joskus jopa **täysin mahdotonta!**

Toisaalta otoskoon kasvaessa perusjoukon systemaattiset piirteet tulevat paremmin esille, eli se vaikuttaa keskeisesti siihen miten hyvin otoksesta tehdyt johtopäätökset voidaan yleistää perusjoukolle.

Onneksi on usein mahdollista määrätä etukäteen otoskoko, jolla tutkimusongelmaan voidaan vastata riittävällä tarkkuudella!

Mitkä asiat vaikuttavat otoskoon määrittämiseen?

Käydään seuraavaksi lyhyesti läpi mitkä asiat vaikuttavat otoskoon määrittämiseen, mutta ohitetaan sen tarkempi käsittely tässä tiivistelmässä.

1. **Perusjoukko**. Tutkimusmuuttujien vaihtelu perusjoukossa vaikuttaa keskeisesti tarvittavaan otoskokoon. Samoin esimerkiksi perusjoukon mahdollinen ryhmärakenne!
2. **Tulosten vaadittu tarkkuus**.
   - 2.1 Kuinka varma halutaan olla, että saadut tulokset yleistyvät perusjoukkoon? Tämä määrittää virhemarginaalin! Suurempi sallittu virhemarginaali tarkoittaa pienempää otoskokoa ja päinvastoin.
   - 2.2 Kuinka varma haluat olla, että otos edustaa joukkoa oikein? Tämä määrittää luottamustason! Suurempi haluttu luottamustaso tarkoittaa suurempaa otoskokoa ja päinvastoin.
3. **Odotetun vastauskadon vaikutus.** Koskee kyselytutkimuksia, joissa usein käy niin että osa kyselytutkimukseen valituista jättää vastaamatta.