TURUN YLIOPISTO

Todennäköisyyslaskennan suppenemiskäsitteet

Vilma Halava

LuK-tutkielma
Toukokuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
FT Pekka Nieminen

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

LuK-tutkielma
Pääaine: Tilastotiede
Tekijä: Vilma Halava
Otsikko: Todennäköisyyslaskennan suppenemiskäsitteet
Ohjaaja: FT Pekka Nieminen
Sivumäärä: 23 sivua
Aika: Toukokuu 2026

Tässä tutkielmassa käsitellään todennäköisyyslaskennan suppenemiskäsitteitä ja niiden käyttöä tilastotieteessä. Aluksi esitellään neljä eri suppenemiskäsitettä: jakaumasuppeneminen, stokastinen suppeneminen, keskineliösuppeneminen ja melkein varma suppeneminen. Todistetaan lisäksi suppenemiskäsitteiden väliset yhteydet. Lopuksi näiden käsitteiden soveltamista tilastotieteessä esitellään sovelluskohteilla, esimerkiksi asymptoottinen normaalisuus, estimaattorijonon tarkentuvuus ja binomijakauman Poisson-approksimaatio.

Asiasanat: suppeneminen, suurten lukujen laki, keskeinen raja-arvolause

Sisällys

1 Johdanto

Todennäköisyyslaskennassa ja tilastotieteessä tarkastellaan usein satunnaismuuttujien jonoja, jotka kuvaavat esimerkiksi havainnosta laskettavia otossuureita otoskoon kasvaessa. Näiden jonojen käyttäytymistä tutkitaan erilaisten suppenemiskäsitteiden avulla, jotka ovat keskeisiä myös matemaattisessa tilastotieteessä. Suppenemiskäsitteet ovat tilastotieteessä suuressa asemassa esimerkiksi estimaattorien tarkentuvuuden tarkastelussa ja estimaattorien asymptoottisen normaaliuden tarkastelussa.

Niiden avulla esitellään myös kaksi matemaattisessa tilastotieteessä merkittävää lausetta, suurten lukujen laki ja keskeinen raja-arvolause. Suurten lukujen laki kuvaa keskiarvojen käyttäytymistä otoskoon kasvaessa ja keskeinen raja-arvolause selittää, miksi monissa tilanteissa otossuureiden jakauma lähestyy normaalijakaumaa.

Tämän työn tavoitteena on esitellä keskeiset satunnaismuuttujien suppenemislajit sekä tarkastella niiden välisiä yhteyksiä. Lisäksi tarkastellaan suurten lukujen lakia ja keskeistä raja-arvolausetta sekä niiden merkitystä tilastollisessa päättelyssä.

Tutkielmassa käytetään standardimerkintöjä tilastotieteen peruskursseilta. Satunnaismuuttujia merkitään $X_1, X_2, ...$, ja satunnaismuuttujien jonoa (X_n), missä indeksi n ilmaisee jonon järjestysluvun eli kuinka mones satunnaismuuttuja jonossa on. Lisäksi käytetään otosavaruudelle merkintää Ω. Otosavaruus on perusjoukko eli kaikki mahdolliset alkeistapaukset, ja merkitään perusjoukkoon kuuluvaa alkiota eli alkeistapausta ω.

Tutkielman jaksossa 2 esitellään neljä suppenemislajia sekä niiden väliset yhteydet. Jaksossa 3 tarkastellaan suurten lukujen lakia ja keskeistä raja-arvolausetta. Jaksossa 4 käsitellään näiden tulosten sovelluksia tilastollisessa päättelyssä, erityisesti estimaattorijonojen tarkentuvuuden ja asymptoottisen normaalisuuden näkökulmasta. Jaksossa 5 esitetään vielä lyhyesti kaksi muuta sovellusta.

Työssä käytetään lähteinä erityisesti kirjoja *Statistical Inference* [1], *Modern Mathematical Statistics* [4] sekä *A Probability Path* [8]. Lisäksi lauseiden todistuksissa hyödynnetään paljon Wikipedia-sivua *Proofs of convergence of random variables* [5].

2 Suppeneminen

Suppenemisella todennäköisyyslaskennassa ja tilastotieteessä viitataan tilanteisiin, joissa satunnaismuuttujien jonon käyttäytyminen lähestyy jotakin tiettyä rajaa jonon indeksin kasvaessa eli toisin sanoen silloin, kun jonossa edetään pidemmälle. Toisin sanoen, satunnaismuuttujien jono lähestyy eli suppenee kohti jotakin tiettyä satunnaismuuttujaa tai vakiota. Joissakin tilanteissa suppenemiskäsitteitä voidaan hyödyntää myös epäsuorassa tarkastelussa. Esimerkiksi kiinnostuksen kohteena olevan satunnaismuuttujan ominaisuuksia voidaan arvioida jonon avulla joka lähestyy kyseistä satunnaismuuttujaa, kun satunnaismuuttujaa ei voida havaita suoraan. Tilastotieteessä suppenemiskäsitteillä on keskeinen rooli erityisesti asymptoottisessa tarkastelussa. Niitä hyödynnetään esimerkiksi estimaattorien tarkentuvuuden analysoinnissa, tilastollisten menetelmien ominaisuuksien tutkimisessa sekä joidenkin jakaumien approksimaatioissa. Suppenemiskäsitteiden avulla voidaan selittää, miksi satunnaisuudesta huolimatta tilastolliset menetelmät ovat luotettavia, kun otoskoko on suuri. Ne siis selittävät, miksi satunnaisuudesta huolimatta otoksista lasketut tunnusluvut antavat luotettavaa tietoa populaation ominaisuuksista. Tästä syystä ne ovat keskeinen osa sekä todennäköisyyslaskentaa että matemaattista tilastotiedettä ja tilastollista päättelyä.

Merkitään suppenemista tilastotieteen peruskurssien mukaisesti: Olkoon $(X_n) = (X_1, X_2, ...)$ jokin satunnaismuuttujajono ja X jokin yksittäinen satunnaismuuttuja. Merkitään jonon (X_n) suppenemista kohti satunnaismuuttujaa X tällöin

$$X_n \to X.$$

Todennäköisyyslaskennan suppenemiskäsitteet tarjoavat työkalut satunnaisilmiöiden pitkän aikavälin käyttäytymisen tarkasteluun. Esitetään seuraavaksi neljä eri suppenemislajia eli neljä eri tapaa, joilla satunnaismuuttujien jono (X_n) lähestyy satunnaismuuttujaa X. Tässä jaksossa käytetään lähteinä erityisesti kirjaa [1] sekä monistetta *Todennäköisyyslaskennan jatkokurssi* [2]. Suppenemislajien määritelmien yhteydessä esitellään lauseet ja todistukset suppenemislajien välisille yhteyksille.

2.1 Suppenemislajit

Työssä esitettävät neljä suppenemislajia ovat: melkein varma suppeneminen, stokastinen suppeneminen, keskineliösuppeneminen ja jakaumasuppeneminen. Kaikissa määritelmissä käytetään merkintää $\lim_{n\to\infty}$, jolla tarkoitetaan raja-arvoa, kun indeksi n kasvaa rajatta. Määritellään käsitteet kirjan [1], kappaleen 5.5 avulla.

Määritelmä 1. Olkoon $(X_n) = (X_1, X_2, ...)$ jokin satunnaismuuttujajono ja X jokin yksittäinen satunnaismuuttuja. Lisäksi merkitään kertymäfunktiota F_n jonolle (X_n) ja F satunnaismuuttujalle X. Kertymäfunktio kertoo todennäköisyyden, että kyseinen satunnaismuuttuja saa arvon, joka on pienempi tai yhtäsuuri kuin luku x. Siis $F_n(x) = P(X_n \leq x)$, ja $F(x) = P(X \leq x)$.

Jono (X_n) *suppenee jakaumaltaan* (engl. convergences in distribution) kohti muuttujaa X, jos niiden kertymäfunktiot lähestyvät toisiaan indeksin n kasvaessa rajatta, ts.

$$\lim_{n\to\infty} F_n(x) = F(x),$$

kaikissa pisteissä x, joissa kertymäfunktio F on jatkuva. Tätä merkitään

$$X_n \xrightarrow{d} X.$$

Tätä suppenemislajia kutsutaan myös heikoksi suppenemiseksi, koska se kuvaa satunnaismuuttujien jakaumien käyttäytymistä. Jakaumasuppeneminen tarkoittaa siis sitä, että satunnaismuuttujien jakaumat lähestyvät rajajakaumaa, mutta se ei välttämättä tarkoita, että satunnaismuuttujat suppenisivat yksittäisten alkeistapausten kohdalla. Toisin sanoen $X_n(\omega) \to X(\omega)$ ei välttämättä päde kaikilla $\omega \in \Omega$. Voi nimittäin olla, että jono (X_n) ja satunnaismuuttuja X eivät ole edes määriteltyjä samassa todennäköisyysavaruudessa, jolloin $X_n(\omega) \to X(\omega)$ ei päde yhdellekään ω.

Määritelmä 2. Olkoon $(X_n) = (X_1, X_2, ...)$ jokin satunnaismuuttujajono ja X jokin yksittäinen satunnaismuuttuja. Jono (X_n) *suppenee stokastisesti* (engl. convergences in probability) kohti muuttujaa X, jos

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \quad \text{kaikilla } \varepsilon > 0$$

tai yhtäpitävästi

$$\lim_{n\to\infty} P(|X_n - X| \leq \varepsilon) = 1 \quad \text{kaikilla } \varepsilon > 0.$$

Merkitään stokastista suppenemista

$$X_n \xrightarrow{p} X.$$

Stokastinen suppeneminen kuvaa todennäköisyysmassan suppenemista (määritelmä 23.1, monisteessa [2]). Tällä tarkoitetaan, että satunnaismuuttujien jonon arvot ovat yhä todennäköisemmin yhä lähempänä rajamuuttujan arvoa, kun indeksi n kasvaa, eli poikkeamien todennäköisyys pienenee mielivaltaisen pieneksi (määritelmä 1.5.5, monisteessa [3]).

Kun satunnaismuuttujien jonossa indeksi n kasvaa kohti ääretöntä, todennäköisyys sille, että satunnaismuuttujien jono ja yksittäisen satunnaismuuttujan erotuksen itseisarvo on suurempaa kuin edellä määritelty ε, lähestyy nollaa. Yhtäpitävästi todennäköisyys sille, että erotuksen itseisarvo on pienempää tai yhtäsuurta kuin ε, lähestyy lukua yksi.

Todistetaan seuraavaksi ensimmäisten kahden suppenemislajin yhteys. Lauseen todistamiseen tarvitaan lisäksi erästä lemmaa, joka esitellään ennen lausetta ja sen todistusta. Käytetään todistuksen lähteenä Wikipedia-sivua *Proofs of convergence of random variables* [5].

Lemma 1. *Olkoon X ja Y satunnaismuuttujia, x jokin reaaliluku ja $\varepsilon > 0$. Nyt*

$$P(Y \leq x) \leq P(X \leq x + \varepsilon) + P(|Y - X| > \varepsilon).$$

Todistus.

$$\begin{aligned}
P(Y \le x) &= P(Y \le x, X \le x+\varepsilon) + P(Y \le x, X > x+\varepsilon) \\
&\le P(X \le x+\varepsilon) + P(Y - X \le x - X, x - X < -\varepsilon) \\
&\le P(X \le x+\varepsilon) + P(Y - X < -\varepsilon) \\
&\le P(X \le x+\varepsilon) + P(Y - X < -\varepsilon) + P(Y - X > \varepsilon) \\
&= P(X \le x+\varepsilon) + P(|Y - X| > \varepsilon)
\end{aligned}$$

□

Nyt lemman 1 avulla voidaan todistaa stokastisen suppenemisen ja jakaumasuppenemisen välinen implikaatio.

Lause 1. *Olkoon* $(X_n) = (X_1, X_2, ...)$ *satunnaismuuttujajono ja* X *jokin yksittäinen satunnaismuuttuja. Jos* $X_n \xrightarrow{p} X$, *niin* $X_n \xrightarrow{d} X$.

Todistus. Oletetaan stokastinen suppeneminen: $X_n \xrightarrow{p} X$, jonka avulla todistetaan, että jakaumasuppeneminen pätee $P(X_n \le x) \to P(X \le x)$, jos oikeanpuoleinen funktio on jatkuva x :ssä. Voidaan kirjoittaa myös toisin kertymäfunktioiden avulla

$$F_n(x) \to F(x).$$

Jakaumasuppenemisen todistamiseksi on osoitettava, että kertymäfunktioiden jono suppenee kohti satunnaismuuttujan X kertymäfunktiota jokaisessa pisteessä x, jossa $F(x)$ on jatkuva. Olkoon x tällainen piste. Kaikille $\varepsilon > 0$ pätee lemman 1 mukaan

$$F_n(x) \le F(x+\varepsilon) + P(|X_n - X| > \varepsilon).$$

Samoin

$$F(x-\varepsilon) \le F_n(x) + P(|X_n - X| > \varepsilon).$$

Nämä kaksi voidaan yhdistää, jolloin saadaan

$$F(x-\varepsilon) - P(|X_n - X| > \varepsilon) \le F_n(x) \le F(x+\varepsilon) + P(|X_n - X| > \varepsilon).$$

Ja stokastisesta suppenemisesta saadun oletuksen

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \to 0$$

perusteella voidaan supistaa epäyhtälöketjua:

$$F(x-\varepsilon) \le \lim_{n\to\infty} F_n(x) \le F(x+\varepsilon).$$

Oletuksen mukaan kertymäfunktio on jatkuva pisteessä x, jolloin epäyhtälöketjun sekä oikea että vasen puoli lähestyvät kohti $F(x)$, kun $\varepsilon \to 0^+$, ts. $\lim_{\varepsilon\to 0} F(x \pm \varepsilon) = F(x)$. Kuristusperiaatteen [6] avulla nähdään, että myös keskelle jäävä $F_n(x)$ suppenee kohti satunnaismuuttujan X kertymäfunktiota, eli $F_n(x) \to F(x)$, ts.

$$X_n \xrightarrow{d} X.$$

□

Jakaumasuppenemisen ja stokastisen suppenemisen yhteys yksinkertaistuu, kun lähestytään vakiota c. Todistetaan vielä tämä tilastotieteessä useasti käytetty erikoistapaus $x = c$, eli jos $X_n \xrightarrow{d} c$, niin $X_n \xrightarrow{p} c$:

Lause 2. *Olkoon c vakio. Jos $X_n \xrightarrow{d} c$, niin $X_n \xrightarrow{p} c$.*

Todistus. Vakion c (tulkittuna satunnaismuuttujaksi) kertymäfunktio:

$$F(x) = \begin{cases} 0, & \text{kun } x < c \\ 1, & \text{kun } x \geq c, \end{cases}$$

on jatkuva kaikissa pisteissä $x \neq c$. Koska satunnaismuuttujien jono lähestyy vakiota c jakaumasuppenemisen mielessä, niin silloin kaikilla $\varepsilon > 0$ on $\lim_{n\to\infty} F_n(c-\varepsilon) = 0$ ja $\lim_{n\to\infty} F_n(c+\frac{\varepsilon}{2}) = 1$ jolloin:

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n - c| \geq \varepsilon) &= \lim_{n\to\infty} [P(X_n \leq c - \varepsilon) + P(X_n \geq c + \varepsilon)] \\
&= \lim_{n\to\infty} P(X_n \leq c - \varepsilon) + \lim_{n\to\infty} P(X_n \geq c + \varepsilon) \\
&= \lim_{n\to\infty} F_n(c - \varepsilon) + \lim_{n\to\infty} P(X_n \geq c + \varepsilon) \\
&= 0 + \lim_{n\to\infty} P(X_n \geq c + \varepsilon) \\
&\leq \lim_{n\to\infty} P\left(X_n > c + \frac{\varepsilon}{2}\right) = 1 - \lim_{n\to\infty} F_n\left(c + \frac{\varepsilon}{2}\right) = 0
\end{aligned}$$

kaikilla $\varepsilon > 0$. Näin ollen $X_n \xrightarrow{p} c$. □

Edellisten perusteella voidaan siis merkitä

$$X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c.$$

Määritelmä 3. Olkoon $(X_n) = (X_1, X_2, ...)$ jokin satunnaismuuttujajono ja X jokin yksittäinen satunnaismuuttuja. Jono (X_n) *suppenee L^r-mielessä* (engl. convergences in r:th mean) kohti muuttujaa X, jos

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0,$$

jossa $r \geq 1$. Merkitään tätä suppenemista

$$X_n \xrightarrow{L^r} X.$$

Tapausta $r = 2$ kutsutaan *keskineliösuppenemiseksi*. Seuraavan lauseen mukaan keskineliösuppeneminen seuraa suppenemisesta L^r-mielessä mille tahansa $r \geq 2$. Edellisen määritelmän perusteella merkitään keskineliösuppenemista

$$X_n \xrightarrow{L^2} X$$

Keskineliösuppeneminen (engl. convergence in mean square) tai kvadraattinen suppeneminen tarkoittaa sitä, että satunnaismuuttujien jonon ja yksittäisen satunnaismuutujan erotuksen neliön odotusarvo lähestyy nollaa, kun n lähestyy ääretöntä, ts.

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0.$$

Lause 3. *Olkoon* $1 \leq r \leq s$. *Jos* $X_n \xrightarrow{L^s} X$, *niin* $X_n \xrightarrow{L^r} X$

Todistus. Oletetaan lauseessa valittu muuttuja r ja $X_n \xrightarrow{L^s} X$. Hölderin epäyhtälö on tunnetusti (ks. lause 4.7.2 [1]):

$$\mathbb{E}[|UV|] \leq (\mathbb{E}[|U|^p])^{\frac{1}{p}} (\mathbb{E}[|V|^q])^{\frac{1}{q}},$$

jossa $1 < p, q < \infty$, ja $\frac{1}{p} + \frac{1}{q} = 1$. Merkitään nyt $U = |X_n - X|^r$, $V = 1$, $p = \frac{s}{r} > 1$. Saadaan, että $\mathbb{E}[|X_n - X|^r] \leq \mathbb{E}[|X_n - X|^{\frac{1}{p}}]$. Oletuksena oli, että $X_n \xrightarrow{L^s} X$, joten

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^s] = 0.$$

Tästä seuraa siis

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] \leq \lim_{n\to\infty} \mathbb{E}[|X_n - X|^s]^{\frac{1}{p}} = 0.$$

Todistus osoittaa siis, että $X_n \xrightarrow{L^r} X$.

$\square$

Huomautus 1. Jos $X_n \xrightarrow{L^s} X$ ja $s \geq 2$, niin silloin myös $X_n \xrightarrow{L^2} X$. Ts. $X_n \to X$ keskineliömielessä.

Lause 4. *Jos* $X_n \xrightarrow{L^r} X$ *jollain* $r \geq 1$, *niin* $X_n \xrightarrow{p} X$.

Todistus. Kun $r \geq 1$, kaikille $\varepsilon > 0$ pätee

$$P(|X_n - X| \geq \varepsilon) = P(|X_n - X|^r \geq \varepsilon^r).$$

Nyt Markovin epäyhtälöstä (ks. lemma 3.8.3 [1]) saadaan

$$P(|X_n - X| \geq \varepsilon) = P(|X_n - X|^r \geq \varepsilon^r) \leq \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r}.$$

Määritelmästä 3 saadaan

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0,$$

josta voidaan päätellä, että

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0,$$

kaikilla $\varepsilon > 0$, ts. $X_n \xrightarrow{p} X$. $\square$

Määritelmä 4. Olkoon $(X_n) = (X_1, X_2, ...)$ jokin satunnaismuuttujajono ja X jokin yksittäinen satunnaismuuttuja. Jono (X_n) *suppenee melkein varmasti* (engl. convergences almost surely (a.s.)) kohti muuttujaa X, jos

$$P\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Tämä voidaan kirjoittaa lyhyemmin: $P(X_n \to X) = 1$. Merkitään melkein varmaa suppenemista

$$X_n \xrightarrow{a.s.} X.$$

Palautetaan mieleen, että satunnaismuuttuja on reaaliarvoinen funktio, joka on määritelty otosavaruudessa Ω. Oletetaan, että otosavaruus sisältää alkeistapauksia, joita merkitään ω, jolloin $X_n(\omega)$ ja $X(\omega)$ ovat otosavaruudessa määriteltyjä funktioita. Satunnaismuuttujien jono suppenee kohti satunnaismuuttujaa melkein varmasti eli todennäköisyydellä 1, ts. lukuunottamatta alkeistapausten joukkoa ω, joiden todennäköisyys on 0, jos ja vain jos funktio $X_n(\omega)$ suppenee kohti $X(\omega)$ melkein kaikilla ω, jotka kuuluvat otosavaruuteen.

Seuraavan lauseen todistus perustuu Wikipedia-sivuun [5], sekä kirjan [8] kappaleeseen 6.2 ja lauseeseen 6.2.1.

Lause 5. *Jos* $X_n \xrightarrow{a.s.} X$, *niin* $X_n \xrightarrow{p} X$.

Todistus. Oletetaan melkein varma suppeneminen $X_n \xrightarrow{a.s} X$. Tiedetään, että melkein varman suppenemisen voi kirjoittaa myös erotuksen itseisarvona, ts.

$$P(\{\omega \mid |X_n(\omega) - X(\omega)| \to 0) = 1.$$

Nyt halutaan siis todistaa melkein varman suppenemisen, ja stokastisen suppenemisen välinen implikaatio. Stokastinen suppeneminen merkitsi siis sitä, että

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \to 0 \quad \text{kaikilla } \varepsilon > 0.$$

Olkoon $\varepsilon > 0$ mielivaltainen. Määritellään nyt tapahtuma A_n kaikille n niin, että

$$A_n = \bigcup_{m\geq n} \{|X_m - X| > \varepsilon\}.$$

A_n on vähenevä, koska kun indeksi n kasvaa, yhdisteessä on yhä vähemmän joukkoja. Raja-arvon $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ määritelmän mukaan on olemassa rajaluku $N(\omega)$ siten, että

$$|X_n(\omega) - X(\omega)| \leq \varepsilon \quad \text{kaikilla } n \geq N(\omega).$$

ts.

$$\omega \notin \bigcup_{n\geq N(\omega)} \{|X_n - X| > \varepsilon\} = A_{N(\omega)}.$$

Oletuksen nojalla tämä pätee melkein kaikilla ω (eli todennäköisyydellä 1), jolloin tapahtuman

$$\bigcap_{n=1}^{\infty} A_n = \{\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon \quad i.o.\},$$

todennäköisyys on 0. Edellä i.o. on lyhenne sanoista *infinitely often*. Melkein varma suppeneminen siis antaa

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = 0.$$

Koska A_n on vähenevä, saadaan:

$$\lim_{n\to\infty} P(A_n) = 0.$$

Koska

$$\{|X_n - X| > \varepsilon\} \subset A_n,$$

niin

$$P(|X_n - X| > \varepsilon) \leq P(A_n).$$

Raja-arvon perusominaisuudesta saadun $P(A_n) \to 0$ perusteella

$$P(|X_n - X| > \varepsilon) \leq P(A_n) \xrightarrow{n\to\infty} 0.$$

Koska valitsemamme ε oli mielivaltainen, niin $X_n \xrightarrow{p} X$.

$\square$

2.2 Suppenemislajien yhteydet kootusti

Käytetään tästä eteenpäin suppenemiskäsitteistä määritelmissä esitettyjä merkintöjä, ja esitellään niiden keskinäiset yhteydet perustuen lauseisiin 1, 3, 4 ja 5 edellä.

$$\begin{tikzcd} \boxed{X_n \xrightarrow{a.s.} X} \arrow[dr] & & \\ & \boxed{X_n \xrightarrow{p} X} \arrow[r] & \boxed{X_n \xrightarrow{d} X} \\ \boxed{X_n \xrightarrow{L^2} X} \arrow[ur] & & \end{tikzcd}$$

Kaavio havainnollistaa esitettyjen neljän eri suppenemiskäsitteen väliset implikaatiot. Nähdään, että vahvimmat suppenemiskäsitteet ovat melkein varma suppeneminen ja keskineliösuppeneminen, jotka eivät implikoi kuitenkaan toisiaan. Muut kolme suppenemiskäsitettä implikoivat jakaumasuppenemisen, mutta jakaumasuppeneminen ei itse implikoi muita suppenemiskäsitteitä ja on siten heikoin esitetyistä käsitteistä.

Tilastotieteessä esiintyy kuitenkin usein yhteyksiin liittyvä erikoistapaus. Silloin satunnaismuuttujien jono lähestyy vakiota c eikä satunnaismuuttujaa. Tällöin kaavio muuttuu lauseen 2 perusteella:

$$\begin{tikzcd} \boxed{X_n \xrightarrow{a.s.} c} \arrow[dr] & & \\ & \boxed{X_n \xrightarrow{p} c} \arrow[r, leftrightarrow] & \boxed{X_n \xrightarrow{d} c} \\ \boxed{X_n \xrightarrow{L^2} c} \arrow[ur] & & \end{tikzcd}$$

Erikoistapauksessa, kun lähestytään vakiota, lauseen 1 mukaan stokastinen suppeneminen ja jakaumasuppeneminen ovat ekvivalentit.

3 Suurten lukujen laki ja keskeinen raja-arvolause

Edellä esitellyt neljä suppenemiskäsitettä muodostavat pohjaa matemaattiseen tilastotieteeseen. Esitellään seuraavaksi kaksi matemaattisen tilastotieteen lausetta, suurten lukujen laki ja keskeinen raja-arvolause. Käytetään lähteinä kirjoja *Modern Mathematical Statistics* [4] ja sen kappaleita 6.2 ja 6.3 sekä kirjaa *Statistical Inference* [1] ja sen kappaletta 5.5 ja lausetta 5.5.2.

Lause 6 (Suurten lukujen laki SLL (heikko versio))**.** *Oletetaan, että satunnaismuuttujat* $X_1, X_2, \ldots$ *ovat riippumattomia ja samoin jakautuneita ja että niillä on odotusarvo* $\mathbb{E}[X_n] = \mu$ *ja varianssi* $\operatorname{Var}[X_n] = \sigma^2 < \infty$.
Tällöin:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Todistus. Todistetaan väite Markovin epäyhtälön avulla (ks. Lemma 3.8.3 [1]). Merkitään

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ja osoitetaan, että kaikilla $\varepsilon > 0$

$$\lim_{n\to\infty} P(\left|\bar{X}_n - \mu\right| < \varepsilon) = 1,$$

mikä siis tarkoittaa, että $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$.

Tiedetään, että jokaiselle $\varepsilon > 0$

$$P(\left|\bar{X}_n - \mu\right| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\varepsilon^2} = \frac{\operatorname{Var}[\bar{X}]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

jolloin

$$P(\left|\bar{X}_n - \mu\right| < \varepsilon) = 1 - P(\left|\bar{X}_n - \mu\right| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{(n\varepsilon^2)} \xrightarrow{n\to\infty} 1.$$

□

Huomautus 2. Edellä oleva todistus osoittaa, että myös:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{L^2} \mu.$$

Huomautus 3. Riippumattomuusoletuksen sijaan edellisessä lauseessa riittäisi myös oletus korreloimattomuudesta eli että

$$\operatorname{Cov}(X_i, X_j) = 0, i \neq j.$$

Suurten lukujen lain perusperiaatteena on selittää, miksi satunnaiset ilmiöt alkavat näyttää säännöllisiltä, kun niitä tarkastellaan riittävän pitkään. Otetaan esimerkkinä kolikonheitto. Jos heität kolikkoa esimerkiksi kuusi kertaa, on mahdollista, että saat viisi kertaa kruunan ja yhden kerran klaavan. Tämä on melko kaukana odotetusta tuloksesta, eli siitä, että molempia olisi noin puolet. Suurten lukujen laki kertoo, että jos jatkaisimme kolikonheittoa todella monta kertaa, kruunan ja klaavan osuudet tasoittuisivat kohti puolikasta. Sanotaan siis, että satunnaismuuttujien jonon keskiarvo suppenee kohti satunnaismuuttujien odotusarvoa. Toisin sanoen heittojen keskiarvo toistojen kasvaessa lähestyy kohti pitkän aikavälin todennäköisintä arvoa. Suurten lukujen laki ei siis kerro, että jos heittäisimme monta kertaa putkeen klaavaa, seuraava olisi välttämättä kruuna, vaan, että jos heitämme tarpeeksi monta kertaa, heittojen jakauma alkaisi tasoittua. Tilastotieteessä suurten lukujen laki antaa perustan otantamenetelmille. Kun otos on riittävän suuri ja satunnaisesti valittu, sen keskiarvo suppenee kohti koko populaation todellista keskiarvoa. Suurten lukujen laki siis kuvaa, miten satunnaisuus muuttuu järjestykseksi, kun havaintoja kertyy paljon. Se selittää, miksi pitkän aikavälin mittaukset ovat luotettavampia kuin yksittäiset havainnot. Tämän vuoksi suurten lukujen lakia voi kutsua matemaattisen tilastotieteen kulmakiveksi.

Lause 7 (Keskeinen raja-arvolause (KRL))**.** *Oletetaan, että satunnaismuuttujat* $X_1, X_2, ...$ *ovat riippumattomia ja samoin jakautuneita, ja että odotusarvo* $\mathbb{E}[X_n] = \mu$ *ja varianssi* $\mathrm{Var}[X_n] = \sigma^2 < \infty$.
Tällöin:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} Z, \qquad \textit{missä } Z \sim N(0, \sigma^2).$$

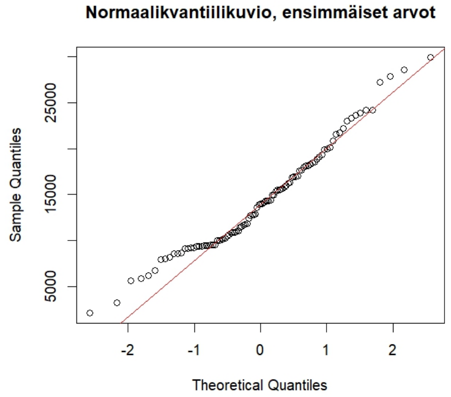
Todistus. Sivuutetaan. Lauseen todistus on esimerkiksi Dudewiczin ja Mishran kirjassa [4], kappaleessa 6.3. □

Keskeinen raja-arvolause on toinen matemaattisen tilastotieteen kulmakivistä. Siinä, missä suurten lukujen laki puhuu keskiarvon asettumisesta oikeaan kohtaan, keskeinen raja-arvolause kertoo, miltä tämän keskiarvon vaihtelu yleensä näyttää. Sen ydinajatus on, että kun otetaan suuri joukko toisistaan riippumattomia havaintoja, niiden summa tai keskiarvo jakautuu likimain normaalisti eli alkaa noudattamaan normaalijakaumaa riippumatta siitä, millainen alkuperäinen jakauma oli. Tämä mahdollistaa normaalijakauman käytön hyvin monissa käytännön tilanteissa, vaikka aineisto ei itsessään olisi normaalisti jakautunutta. Esimerkkinä voidaan ajatella ihmisten päivittäistä askelten määrää. Toiset kävelevät vähän ja toiset paljon, jakauma on epäsymmetrinen ja pitkähäntäinen. Mutta jos tarkastellaan esimerkiksi kymmenen päivän keskiarvoa sadalta ihmiseltä, keskiarvot alkavat muistuttaa normaalijakaumaa. Yksittäisten päivien vaihtelu niin sanotusti kesyyntyy, kun ne yhdistetään. Toisena esimerkkinä, jos kolikkoa heitetään erittäin monta kertaa, todennäköisyys saada kruuna tietyllä määrällä heittoja lähestyy normaalijakaumaa.

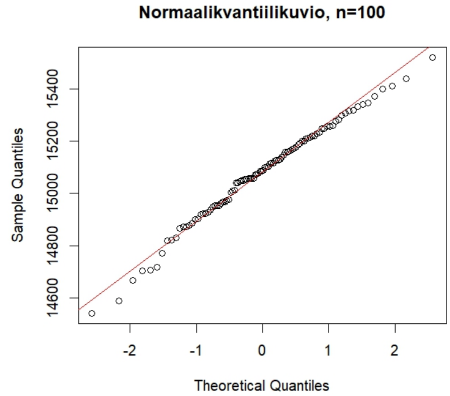
Keskeinen raja-arvolause on se syy, miksi normaalijakaumaa käytetään lähes kaikkialla: mittausvirheiden mallintamisessa, kyselytutkimusten virhemarginaalien laskemisessa, kokeiden analyysissa ja monissa muissa tilastollisissa menetelmissä. Se tarjoaa perustan tilastollisille työkaluille, joilla arvioidaan, onko havaittu ero todellinen vai pelkkää satunnaisvaihtelua. Keskeinen raja-arvolause siis kertoo, että vaikka maailma on monimutkainen, suurissa kokonaisuuksissa on yllättävän paljon säännöllisyyttä.

Seuraavaksi esimerkki keskeisestä raja-arvolauseesta:

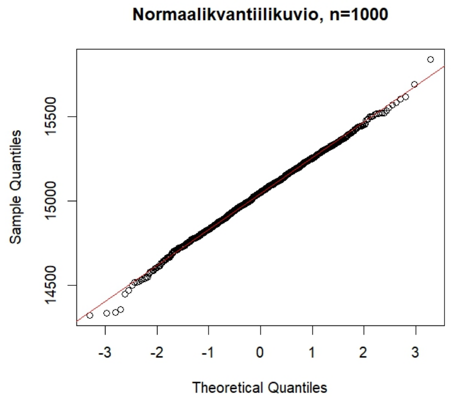
Esimerkki 1. Tarkastellaan Turun yliopiston aktiivisuustutkimuksen aineistoa, jossa on kerätty sadan ihmisen askeldataa noin kahdeksan päivän ajan. Aineistossa askeldataa on vain noin 750 päivältä, sillä aineistosta on poistettu päivät, jolloin askelmääriä laskevaa aktiivisuusranneketta on käytetty alle kymmenen tuntia päivässä. Tästä syystä joillain tutkittavilla aineistossa askelmäärää on vähemmiltä päiviltä kuin toisilla. Tässä työssä käytettävässä aineistossa pidettiin vain kaksi saraketta, tutkittavien id:t ja päiväkohtaiset askelmäärät. Muut aktiivisuusrannekkeen keräämät tiedot poistettiin. Tehdään normaalikvantiilikuvio jokaisen aineistossa olevan tutkittavan yksilön ensimmäisestä askeldatasta, ts. ensimmäisestä aktiivisuusrannekeenkäyttöpäivästä. Koska aineisto on pieni, käytetään tässä keskeisen raja-arvolauseen esimerkissä hyödyksi bootstrapping-menetelmää. Bootstrappingissä muodostetaan suuri määrä (esim. 10 000) satunnaisotoksia alkuperäisestä aineistosta takaisinpanolla. Jokaisesta otoksesta lasketaan askelmäärien keskiarvo, näin saadaan jakauma keskiarvoista.



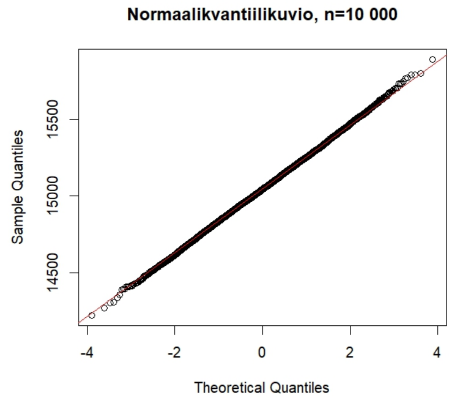
(a) Ensimmäisen mittauspäivän askelmäärät



(b) $n = 100$



(c) $n = 1000$



(d) $n = 10000$

Kuva 1: Normaalikvantiilikuviot

Kuvassa 1 esitetään normaalikvanttiilikuviot (QQ-kuviot) bootstrapping-menetelmällä saaduista keskiarvoista eri otosmäärillä (n = 100, 1000 ja 10 000) sekä alkuperäisen aineiston ensimmäisistä havainnoista. QQ-kuvioiden avulla tarkastellaan, kuinka hyvin aineisto noudattaa normaalijakaumaa: mitä lähempänä pisteet ovat suoraa viivaa, sitä paremmin normaalijakauma kuvaa aineistoa. Kuvasta 1a nähdään, että alkuperäisen aineiston ensimmäiset arvot jokaiselta henkilöltä poikkeaa jonkin verran normaalijakaumasta, mikä ilmenee pisteiden kaareutumisena erityisesti jakauman ääripäissä. Tämä viittaa siihen, että alkuperäinen aineisto ei ole täysin normaalijakautunut.

Kuvissa 1b, 1c, 1d esitetään bootstrapping-menetelmällä muodostettujen otoskeskiarvojen QQ-kuviot kasvavilla otosmäärillä. Kun otosten määrä on 100, pisteet asettuvat jo melko lähelle suoraa, mutta pieniä poikkeamia esiintyy edelleen. Otosmäärän kasvaessa 1000:een ja edelleen 10 000:een pisteet asettuvat yhä tarkemmin suoralle viivalle.

Vaikka alkuperäinen aineisto ei ole normaalijakautunut, otoskeskiarvojen jakauma lähestyy normaalijakaumaa, kun otosten määrä kasvaa. Erityisesti kuvassa 1d jakauma on jo erittäin lähellä normaalijakaumaa, mikä näkyy lähes täydellisen lineaarisena QQ-kuviona. Esimerkissä tapahtuu siis juuri niin, miten keskeinen raja-arvolause kertoo, että tapahtuu. Otoskoon kasvaessa jakauma alkaa noudattaa normaalijakaumaa riippumatta alkuperäisestä jakaumasta.

Normaalijakautuneisuuden tarkastamiseksi voidaan lisäksi tarkastella otoksia myös Shapiro-Wilk-testillä (lukuunottamatta otosta jossa n=10 000, sillä Shapiro-Wilk-testin isoin mahdollinen n on 5000). Shapiro-Wilk-testin mukaan otos on normaalijakautunut, jos p-arvo on yli 0.05.

```
shapiro.test(eka_arvo$steps_total)

Shapiro-Wilk normality test

data:  eka_arvo$steps_total
W = 0.96785, p-value = 0.01594

> shapiro.test(z)

Shapiro-Wilk normality test

data:  z
W = 0.97688, p-value = 0.07575

> shapiro.test(y)

Shapiro-Wilk normality test

data:  y
W = 0.99845, p-value = 0.5247
```

Ensimmäisten arvojen Shapiro-Wilk-testi tehty ensin, sen jälkeen tehty n=100 (aineisto `z`) ja viimeisenä n=1000 (aineisto `y`). Pelkkien ensimmäisten arvojen aineiston Shapiro-Wilk-testi antaa p-arvoksi 0.01594, joka on alle 0.05, eli ensimmäisten arvojen jakauman normaalisuutta vastaan hieman näyttöä. Kun n=100 ja n=1000, Shapiro-Wilk-testi antaa molemmissa tapauksissa p-arvon, joka on yli 0.05. Nämä tulokset (0.07575 ja 0.5247) eivät anna näyttöä normaalisuutta vastaan. Shapiro-Wilk-testi siis puoltaa normaalikvantiilikuvioista saatua tulosta, eli keskiarvotetut otokset aineistosta alkavat otoskoon kasvaessa noudattamaan yhä enemmän ja enemmän normaalijakaumaa.

4 Sovelluskohteita tilastollisessa päättelyssä

Todennäköisyyslaskennan suppenemiskäsitteillä on sovelluskohteita myös tilastollisessa päättelyssä. Tilastollisessa päättelyssä keskeisenä tavoitteena on tehdä johtopäätöksiä tuntemattomista populaation parametreista otoksen perusteella. Koska käytettävissä on vain rajallinen määrä havaintoja, parametreja joudutaan arvioimaan satunnaismuuttujien avulla, joita kutsutaan estimaattoreiksi. Siis jos $X_1, ..., X_n$ ovat havaintoja edustavia satunnaismuuttujia ja θ estimoitava parametri, estimaattori määritellään usein tilastotieteessä havaintojen funktiona:

$$\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n).$$

Estimaattori on siis satunnaismuuttuja ja sen havaittu arvo estimaatti. Tilastollinen estimointi tarkoittaa siis käytännössä oikean mallin valitsemista havaintojen perusteella ([8], kappale 6.2.1).

Suppenemiskäsitteet tarjoavat teoreettisen perustan sille, miksi tällainen päättely on mahdollista. Jos estimaattorijono suppenee esimerkiksi stokastisesti kohti tarkasteltavaa parametria, kun havaintojen lukumäärä n kasvaa rajatta, estimaattoria kutsutaan tarkentuvaksi. Tällöin estimaattorin arvot lähestyvät todellista parametria otoskoon kasvaessa. Lisäksi monet tilastolliset menetelmät perustuvat siihen, että estimaattorien jakaumaa voidaan approksimoida normaalijakaumalla suurilla otoskoilla. Tämä seuraa keskeisestä raja-arvolauseesta ja estimaattorien asymptoottisesta normaalisuudesta. Näiden tulosten avulla voidaan esimerkiksi muodostaa luottamusvälejä ja suorittaa hypoteesitestejä myös tilanteissa, joissa tarkkaa jakaumaa ei tunneta. Esitellään seuraavaksi kaksi sovelluskohdetta tilastollisessa päättelyssä, estimaattorijonon tarkentuvuus (määritelmä 5) ja suurimman uskottavuuden estimaatin asymptoottinen normaalisuus (määritelmä 6).

4.1 Estimaattorijonon tarkentuvuus

Tarkastellaan tilastollista mallia $f_{\mathbf{X_n}}(\mathbf{x}; \theta)$ aineistolle $\mathbf{X}_n = (X_1, ..., X_n)$. Olkoon θ parametri ja

$$\hat{\theta}_n = \hat{\theta}(X_1, ..., X_n)$$

sille muodostettu estimaattori.

Määritelmä 5. Estimaattorijono $\hat{\theta}_n$ on *tarkentuva*, jos $\hat{\theta}_n \xrightarrow{p} \theta$. Ts. estimaattorijono on tarkentuva, jos estimaattorijono suppenee stokastisesti kohti parametrin todellista arvoa.

Esimerkki 2. Oletetaan, että $X_1, ..., X_n$ ovat riippumattomia satunnaismuuttujia jostakin seuraavista kolmesta jakaumasta, joissa kaikissa on odotusarvo μ: normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, Poisson-jakaumasta $\mathrm{Poisson}(\mu)$ tai eksponenttijakaumasta $\mathrm{Exp}(\frac{1}{\mu})$. Koska kaikissa on sama odotusarvo, voidaan tarkastella jakaumia yhtä aikaa, ja merkitään jakaumia seuraavanlaisesti; $X_1, ..., X_n \sim \mathrm{N}(\mu, \sigma^2)$, $X_1, ..., X_n \sim \mathrm{Poisson}(\mu)$, $X_1, ..., X_n \sim \mathrm{Exp}(\frac{1}{\mu})$. Käytetään odotusarvon estimaattorina otoskeskiarvoa:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n.$$

Tarkastellaan, onko otoskeskiarvo odotusarvon tarkentuva estimaattori, toisin sanoen, päteekö

$$\hat{\mu}_n \xrightarrow{p} \mu.$$

Käytetään apuna suurten lukujen lakia. Suurten lukujen lain mukaan, jos odotusarvo on μ ja varianssi $< \infty$, niin silloin

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Nyt jokaisessa tarkasteltavassa jakaumassa odotusarvona on μ ja tiedetään jakaumista, että varianssi noudattaa myös annettua ehtoa. Tällöin suurten lukujen lain mukaan otoskeskiarvo suppenee stokastisesti kohti jakauman odotusarvoa. Ts.

$$\hat{\mu}_n \xrightarrow{p} \mu.$$

Käytetään seuraavan esimerkin lähteenä Stanfordin yliopiston *Properties of Estimators* opetusdioja [7].

Esimerkki 3. Oletetaan, että satunnaismuuttujat $X_1, ..., X_n$ ovat riippumattomia ja noudattavat välin $[0, \theta]$ tasajakaumaa, ts.

$$X_1, ..., X_n \sim \mathrm{Tas}(0, \theta),$$

jossa $\theta > 0$ on tuntematon parametri. Tasajakaumassa satunnaismuuttujan tiheys on vakio välillä $[0, \theta]$, ja satunnaismuuttujan arvo ei voi ylittää parametria θ. Parametri θ on siis jakauman yläraja. Satunnaismuuttujien tiheysfunktio on siis:

$$f(x) = \begin{cases} \frac{1}{\theta}, & \text{kun } 0 \le x \le \theta \\ 0, & \text{muulloin.} \end{cases}$$

Kertymäfunktion määritelmä:

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Tarkastellaan kolmea tapausta: Jos $x < 0$, silloin $f(t) = 0$ kaikilla $t \le x$, jolloin kertymäfunktio $F(x) = 0$. Jos $0 \le x \le \theta$, silloin kertymäfunktio $F(x) = \int_0^x \frac{1}{\theta}dt = \frac{x}{\theta}$. Ja jos $x > \theta$, jolloin integraali otetaan yli koko alueen, jolloin kertymäfunktio on $F(x) = 1$. Täten kaava kertymäfunktiolle on:

$$F(x) = \begin{cases} 0, & \text{kun } x < 0 \\ \frac{x}{\theta}, & \text{kun } 0 \le x \le \theta \\ 1, & \text{jos } x > \theta \end{cases}.$$

Tarkastellaan tasajakauman parametrin estimaattoria, joka on otosmaksimi, ts.

$$\hat{\theta}_n = \max(X_1, ..., X_n).$$

Kaikille $x \in \mathbb{R}$:

$$P(\hat{\theta}_n \leq x) = P(X_1 \leq x, ..., X_n \leq x).$$

Lisäksi, koska satunnaismuuttujat ovat riippumattomia, pätee

$$P(X_1 \leq x, ..., X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x).$$

Ja koska satunnaismuuttujat ovat samoin jakautuneita, pätee: $P(X_i \leq x) = F(x)$, jolloin

$$P(\hat{\theta}_n \leq x) = (F(x))^n.$$

Tarkastellaan, onko suurimman uskottavuuden estimaattori $\hat{\theta}_n$ tarkentuva. Määritelmän 5 mukaan estimaattori on tarkentuva, jos

$$\forall \varepsilon > 0 : P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \xrightarrow{n \to \infty} 0.$$

Voidaan merkitä:

$$P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = P\left(\hat{\theta}_n > \theta + \varepsilon\right) + P\left(\hat{\theta}_n < \theta - \varepsilon\right),$$

josta nähdään, että meillä on kaksi vaihtoehtoa saada tarkentuvuuden määritelmästä itseisarvo suuremmaksi kuin ε. Jos $\varepsilon > 0$, silloin $\theta + \varepsilon > \theta$ ja $F(\theta + \varepsilon) = 1$, jolloin $P(\hat{\theta}_n \leq \theta + \varepsilon) = (F(\theta + \varepsilon))^n = 1$, josta seuraa, että $P(\hat{\theta}_n > \theta + \varepsilon) = 0$. Jäljelle jää siis vain:

$$P(\hat{\theta}_n < \theta - \varepsilon) = P(\hat{\theta}_n \leq \theta - \varepsilon) = (F(\theta - \varepsilon))^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 0,$$

koska

$$0 < \frac{\theta - \varepsilon}{\theta} < 1.$$

Ts.

$$\lim_{n \to \infty} P(\hat{\theta}_n \leq \theta - \varepsilon) = 0.$$

Siis, kun indeksi n kasvaa rajatta, on otosmaksimi tarkentuva estimaattori parametrille θ, eli estimaattori suppenee stokastisesti kohti parametrin todellista arvoa:

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

4.2 Asymptoottinen normaalisuus

Määritelmä 6. Olkoon $\hat{\theta}_n$ estimaattori parametrille θ. Estimaattori $\hat{\theta}_n$ on *asymptoottisesti normaalinen*, jos:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z,$$

kun $Z \sim N(0, v(\theta))$ ja $v(\theta) > 0$ on jokin thetasta riippuva lauseke.

Asymptoottisella normaalisuudella tarkoitetaan tilannetta, jossa estimaatin jakauma lähestyy eli suppenee kohti normaalijakaumaa otoskoon n kasvaessa rajatta. Käytännössä tämä tarkoittaa sitä, että kun otoskoko on riittävän suuri estimaatin jakaumaa voidaan approksimoida normaalijakaumalla. Asymptoottinen normaalisuus on keskeinen tulos tilastollisessa päättelyssä, sillä sen avulla voidaan muodostaa esimerkiksi luottamusvälejä, tehdä approksimaatiotia ja suorittaa hypoteesitestejä. Asymptoottinen normaalisuus perustuu edellä esitettyyn keskeiseen raja-arvolauseeseen. Yleisesti säännöllisyysehtojen vallitessa suurimman uskottavuuden estimaattorit ovat asymptoottisesti normaalisia [7]. Tiedetään siis, että mitä suurempi otoskoko, sitä paremmin suurimman uskottavuuden estimaatin jakauma noudattaa normaalijakaumaa.

Johdetaan estimaattorin asymptoottisen normaalisuuden avulla kaava likimääräiselle luottamusvälille seuraavassa esimerkissä kirjan [9] kappaleen 9.7 avulla.

Esimerkki 4. Lähtökohtana asymptoottinen normaalisuus, ts. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z$, missä $Z \sim N(0, v(\theta))$. Voidaan merkitä

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, v(\theta)).$$

Nyt standardoinnilla saadaan

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{v(\theta)}} \xrightarrow{d} N(0, 1),$$

ja kun n on iso,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{v(\theta)}} \approx Z, \quad Z \sim N(0, 1).$$

Merkitään standardinormaalijakauman $(1 - \alpha/2)$-kvantiilia $z_{\alpha/2}$, jolloin

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Käytetään yllä olevaa Z arviota, ja kirjoitetaan approksimaatio:

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{v(\theta)}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Muokkaamalla ja uudelleenkirjoittamalla saadaan:

$$P\left(\frac{-z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}} \leq \hat{\theta}_n - \theta \leq -\frac{z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}}\right) \approx 1 - \alpha$$

$$P\left(\hat{\theta}_n - \frac{z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + \frac{z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Täten, likimääräinen $(1 - \alpha)$ luottamusväli on:

$$\left[\hat{\theta}_n - \frac{z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}}, \hat{\theta}_n + \frac{z_{\alpha/2}\sqrt{v(\theta)}}{\sqrt{n}}\right].$$

Käytännössä varianssifunktio $v(\theta)$ on yleensä tuntematon, jolloin edellinen luottamusväli ei ole suoraan laskettavissa. Tämän vuoksi parametrin θ tilalle korvataan estimaattori $\hat{\theta}_n$, jolloin saadaan estimaatti $v(\hat{\theta}_n)$. Kun funktio v on jatkuva, voidaan osoittaa, että

$$v(\hat{\theta}_n) \xrightarrow{p} v(\theta).$$

Tällöin asymptoottinen normaalisuus säilyy myös silloin, kun tuntematon varianssifunktio korvataan estimaatillaan. Tämä perustuu Slutskyn lauseeseen (ks. lause 5.5.15 [1]). Saadaan siis lopullinen muoto:

$$\left[\hat{\theta}_n - \frac{z_{\alpha/2}\sqrt{v(\hat{\theta}_n)}}{\sqrt{n}}, \hat{\theta}_n + \frac{z_{\alpha/2}\sqrt{v(\hat{\theta}_n)}}{\sqrt{n}}\right].$$

5 Muita jakaumasuppenemisen sovelluksia

Esitellään lopuksi lyhyesti vielä kaksi jakaumasuppenemista hyödyntävää tulosta, jotta pystytään havainnoimaan sitä, kuinka laajalti suppenemiskäsitteitä käytetään ja miten erilaisissa sovelluskohteissa.

Lause 8. *Binomijakauman Poisson-approksimaatio:*
Oletetaan, että

$$X_n \sim \mathrm{Bin}(n, p_n), \quad n = 1, 2, ...,$$

jossa $p_n \to 0$*. Jos otoskoko eli indeksi* n *kasvaa rajatta ja*

$$\lim_{n\to\infty} np_n = \lambda \in (0, \infty)$$

niin silloin:

$$X_n \xrightarrow{d} X \sim \mathrm{Poisson}(\lambda),$$

eli binomijakauma lähestyy Poisson-jakaumaa.

Lauseen 8 todistuksessa seurataan kirjan [10] lauseen 4.8.3 todistusta:

Todistus. Todistetaan lauseen väite, jossa $n \to \infty$, $p_n \to 0$, jolloin $np_n \to \lambda$, jossa $\lambda > 0$ on kiinteä.

Binomijakauman pistetodennäköisyysfunktio on

$$P(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}.$$

Uudelleenkirjoittamalla binomikerroin muotoon $\frac{n!}{k!(n-k)!}$ saadaan pistetodennäköisyysfunktioksi:

$$P(X_n = k) = \frac{n!}{k!(n-k)!} p_n^k \left(1 - p_n\right)^n \left(1 - p_n\right)^{-k}.$$

Ensimmäinen termi voidaan merkitä vielä toisin, kun jaetaan ensin keskenään $n!$ ja $(n-k)!$. Lisäksi voidaan avata sulkeet toisesta termistä ja muokata pistetodennäköisyysfunktiota:

$$\begin{aligned}
P(X_n = k) &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(p_n\right)^k \left(1 - p_n\right)^n \left(1 - p_n\right)^{-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{(np_n)^k}{n^k} \left(1 - p_n\right)^n \left(1 - p_n\right)^{-k} \\
&= \frac{(np_n)^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - p_n\right)^n \left(1 - p_n\right)^{-k}
\end{aligned}$$

Tarkastellaan pistetodennäköisyysfunktion termejä erikseen, kun $n \to \infty$ ja k on kiinteä:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{n(n-1)...(n-k+1)}{n^k} \\
&= \lim_{n\to\infty} \left(\frac{n}{n} \times \frac{n-1}{n} \times ... \times \frac{n-k+1}{n}\right) \\
&= \lim_{n\to\infty} \left(1 \times \left(1 - \frac{1}{n}\right) \times ... \times \left(1 - \frac{k-1}{n}\right)\right) \\
&= 1,
\end{aligned}$$

sillä $1 - \frac{i}{n} \to 1$.

Koska oletuksen mukaan $np_n \to \lambda$, saadaan:

$$(np_n)^k \to \lambda^k.$$

Lisäksi kirjan [10] kaavan A.2.5 perusteella saadaan termille $(1-p_n)^n$:

$$(1-p_n)^n \xrightarrow{n\to\infty} \mathrm{e}^{-\lambda}.$$

Koska $p_n \to 0$ ja k on kiinteä, saadaan viimeiselle termille:

$$(1-p_n)^{-k} \xrightarrow{n\to\infty} 1.$$

Yhdistämällä saadut tulokset saadaan

$$P(X_n = k) \to \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},$$

kun $n \to \infty$. Koska tämä on Poisson(λ)-jakauman pistetodennäköisyysfunktio, seuraa:

$$X_n \xrightarrow{d} \text{Poisson}(\lambda).$$

□

Tämä tulos on hyödyllinen tilastollisessa päättelyssä, sillä se mahdollistaa binomikokeiden approksimoimisen Poisson-jakaumalla suurilla otoskooilla ja pienillä onnistumisen todennäköisyyksillä. Tämä siis helpottaa esimerkiksi todennäköisyyksien laskemista harvinaisten tapahtumien yhteydessä.

Palautetaan seuraavaksi mieleen t_n-jakauman määritelmä. Jos

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

ja

$$V_n \sim \chi^2_n,$$

missä satunnaismuuttujat Z ja V_n ovat riippumattomia, niin satunnaismuuttuja

$$T_n = \frac{Z}{\sqrt{V_n/n}}$$

noudattaa t-jakaumaa t_n, missä n on jakauman vapausasteiden määrä.

Lause 9. *Kun vapausasteet kasvavat rajatta, t-jakauma lähestyy kohti standardinormaalijakaumaa. Oletetaan, että $T_n \sim t_n$, jossa $n = 1, 2, ...$, on vapausasteiden määrä. Nyt, kun $n \to \infty$, pätee*

$$T_n \xrightarrow{d} Z \sim \mathrm{N}(0,1),$$

eli t-jakauma suppenee jakaumaltaan kohti standardinormaalijakaumaa.

Lauseen 9 todistuksen lähteenä käytetty kirjan [10] lauseen 10.4.5 todistusta ja Slutskyn lausetta [1].

Todistus. Oletetaan siis, että $T_n \sim t_n$. Tiedetään, että

$$T_n = \frac{Z}{\sqrt{\frac{V_n}{n}}},$$

jossa $Z \sim \mathrm{N}(0,1), V_n \sim \chi^2_n$,ja Z sekä V_n keskenään riippumattomia.
Koska $V_n = \sum_{i=1}^n Z_i^2$, jossa $Z \sim \mathrm{N}(0,1)$, niin

$$\mathbb{E}[Z_i^2] = 1.$$

Suurten lukujen laista saadaan:$\frac{V_n}{n} \xrightarrow{p} 1$, jolloin myös $\sqrt{\frac{V_n}{n}} \xrightarrow{p} 1$. Ja Slutskyn lauseella saadaan:

$$T_n = \frac{Z}{\sqrt{\frac{V_n}{n}}} \xrightarrow{d} Z \sim \mathrm{N}(0,1).$$

$\square$

Suuren otoskoon tapauksessa t-jakaumaa voidaan approksimoida normaalijakaumalla. Tämä helpottaa luottamusvälien muodostamista ja hypoteesitestien suorittamista, sillä normaalijakaumaan perustuvat menetelmät antavat silloin tarkkoja ja luotettavia tuloksia.

Viitteet

[1] G. Casella & R. L. Berger: *Statistical Inference*, 2. laitos, Duxbury Press, 2002

[2] J. Lempa & H. Saarinen: *Todennäköisyyslaskennan jatkokurssi*, Turun yliopisto

[3] H. Nyberg, P. Nieminen: *Tilastollinen päättely I & II*, Turun yliopisto, syksy 2024

[4] E. J. Dudewicz & S. N. Mishra: *Modern Mathematical Statistics*, Wiley, 1988

[5] Wikipedia: *Proofs of convergence of random variables*, luettu 28.2.2026, Linkki: Proofs of convergence of random variables

[6] Wikipedia: *Kuristusperiaate*, luettu 2.3.2026, Linkki: Kuristusperiaate

[7] Alex Tsun: *Statistical Estimation, Properties of Estimators II*, Linkki: Properties of Estimators II, Stanfordin yliopiston opetusdiat

[8] Sidney I. Resnick: *A Probability Path*, 5. painos, Birkhauser Boston 2005

[9] Larry A. Wasserman: *All of Statistics*, Springer 2004

[10] Joseph K. Blitzstein & Jessica Hwang: *Introduction to Probability*, 2. laitos, CRC Press, 2019