Rakenteiden Mekaniikka
Vol. 41, Nro. 2, 2008, s. 79 – 85

Ääriarvoista

Esko Valkeila

Tiivistelmä. Artikkelissa esitellään lyhyesti ääriarvojen teorian perusteet sekä niihin liittyvää tilastollista päättelyä.

Avainsanat: ääriarvot, tilastollinen päättely, rakenteiden luotettavuus

Johdanto

Ääriarvo

Ääriarvo on satunnaismuuttujajoukon suurin tai pienin arvo. Koska satunnaismuuttujan $-X$ maksimi on sama kuin satunnaismuuttujan X minimi, niin riittää tarkastella pelkästään satunnaismuuttujajoukon maksimien käyttäytymistä. Mikäli havainnot ovat riippumattomia ja samoin jakautuneita, niin pitkään on ollut jo tunnettua että tällöin maksimilla voi olla vain kolme erilaista rajakaumaa.

Menetelmät ovat luonteeltaan ei-parametrisia. Tämä tarkoittaa sitä, että havaintojen jakaumasta ei tehdä juuri muuta oletusta kuin se, että havaintojen kertymäfunktio on jatkuva. Viime vuosina etenkin tilastollinen päättely on kehittynyt voimakkaasti eteenpäin.

Tässä artikkelissa käsitellään vain reaaliarvoisia [yksiulotteisia] satunnasmuuttujia. Ääriarvojen teoriaa on yleistetty myös satunnaisvektoreille ja stokastisille prosesseille. Arvelen, että juuri satunnaisvektorien ääriarvoihin liittyvät uudet tulokset lisäävät näiden tilastomatemaattisten menetelmien käyttöä tekniikassa.

Ääriarvot ja rakenteiden luotettavuus

Miten ääriarvot liittyvät rakenteiden luotettavuuteen? Rakenteiden luotettavuus riippuu usein juuri isoista kuormituksista, ts. ääriarvoista. Havaittujen ääriarvojen perusteella voi ennustaa jonkun muuttujan suurimman mahdollisen arvon. Menetelmien käyttöä rajoittaa se, että havaintoja tavitaan paljon. Mikäli halutaan mitoittaa rakenteita kestämään kovaa tuulta tai muita sääilmiöihin liittyviä ääriarvoja, niin havaintoja on yleensä tarpeeksi tilastollisen analyysin tekoon. Klassinen esimerkki on merenpinnan maksimikorkeuden ennustaminen, kun käytettävissä on historiallisia havaintoja päivittäisestä maksimikorkeudesta. Patoja rakennettaessa kiinnostaa esimerkiksi tietää, millä todennäköisyydella merenpinnan maksikorkeus voi olla kymmenen prosesenttia suurempi kuin suurin tähän mennessä havaittu korkeus. Joskus taas kiinnostaa selvittää sitä, mikä on periaatteessa suurin mahdollinen arvo, minkä satunnaismuuttuja voi saada. Meitä kaikkia kiinnostava kysymys on se, voidaanko ihmisen maksimaalinen elinikä määrätä. Ääriarvojen teoria tarjoaa yhden vaihtoehdon selvittää tätä.

Esimerkki siitä, miten rakenteiden luotettavuus on vaikuttanut ääriarvojen teoriaan liittyy *heikoimman lenkin* identifiointiin yksitäisten komponenttien minimiarvojen avulla. Jo Leonardo da Vinci arveli että pitkä köysi on heikompi kuin lyhyt köysi ja Galileo Galilei puolestaan osoitti, että näin ei välttämättä ole. Viime vuosisadan alussa muoiltiin edellä mainittu heikoimman lenkin periaate ja yhdistettiin se ääriarvojen tilastolliseen teoriaan, ja voidaankin katsoa että ääriarvojen teoriassa keskeisen Weibullin jakauman synty liittyi juuri näihin tutkimuksiin.

Ääriarvoilla on siis sovelluksia rakenteiden luotettavuutta suunniteltaessa, mutta myös rakenteiden luotettavuuden ongelmat ovat vaikuttaneet ääriarvoihin liittyvän tilastollisen teorian kehitykseen.

Ääriarvojen luokittelu

Tilastomatematiikkaa

Äärivojen teoria on osa tilastomatematiikka. Teorian perusteet on kehitetty jo runsas kuusikymmentä vuotta sitten, mutta vielä äskettäin kehitystä on tapahtunut liittyen moniulotteisiin ääriarvoihin sekä käytännön kannalta tärkeään tilastollisen päättelyn monipuolistumiseen.

Palautetaan mieleen eräitä tilastomatematiikan käsitteitä. Muuttuja, jonka loppuarvoa ei voi ennustaa, voidaan ajatella satunnaismuuttujaksi. Merkitään satunnaismuuttujaa symbolilla X. Satunnaismuuttujan käyttäytymistä analysoidaan sen kertymäfunktion G avulla:

$$G(z) = P(X \leq z);$$

tässä P on todennäköisyys ja kertymäfunktion G arvo pisteessä z kertoo sen, kuinka usein satunnaismuuttujan arvo X on pienempi kuin z: jos satunnaismuuttujasta X saadaan riippumattomia havaintoja 'paljon', niin noin $100 * G(z)$ prosenttia havainnoista on pienempiä kuin z.

Ajatellaan nyt, että teemme satunnaismuuttujsta X n kappaletta riippumattomia havaintoja. Tätä voidaan mallintaa siten, että meillä on satunnaismuuttujat $X_1, \ldots, X_n$ ja ne ovat riippumattomia ja samoin jakautuneita. Suurinta arvoa kuvaa nyt satunnaismuuttuja M_n, missä

$$M_n = \max_{1 \leq k \leq n} X_k.$$

Maksimi on pienempi kuin z, jos kaikki satunnaismuuttujat ovat pienempiä kuin z; tästä saadaan välitömästi, käyttämällä sitä että satunnaismuuttujat X_k ovat riippumattomia ja samoin jakautuneita, että

$$P(M_n \leq z) = (P(X_1 \leq z))^n = G(z)^n. \tag{1}$$

Ääriarvojen rajajakaumien luokittelu

Kertymäfunktiolle pätee, että $G(z) \leq 1$. Käyttämällä tätä tietoa havaitaan helposti, että jos $n \to \infty$, niin $G(z)^n \to 0$ tai $G(z)^n = 1$ kaikilla $n \geq 1$. Mielenkiintoisia raja-arvoja löytyy vain, jos skaalataan ja siirretään satunnaismuuttujaa M_n:

Olkoot b_n, a_n, $a_n > 0$ reaalilukuja ja olkoon

$$M_n^* := \frac{M_n - b_n}{a_n}.$$

Olkoon edelleen G_n satunnaismuuttujan M_n^* kertymäfunktio:

$$G_n(z) := P(M_n^* \leq z) = P(\frac{M_n - b_n}{a_n} \leq z) = G(a_n(z + b_n))^n.$$

Oletetaan, että on olemassa kertymäfunktio F ja lukujonot a_n, b_n siten, että

$$G_n(z) = P(\frac{M_n - b_n}{a_n} \leq z) = G(a_n(z + b_n))^n \to F(z), \tag{2}$$

kun $n \to \infty$.

Voidaan osoittaa, että relaatio $G(a_n(z + b_n))^n \to F(z)$ on yhtäpitävä relaation

$$\lim_{n\to\infty} n\,(1 - G(a_n(z + b_n))) = -\log F(z) \tag{3}$$

kanssa.

Jos oletaan, että on olemassa vakiot b_n, a_n siten että $G_n(z) \to F(z)$ ja F on kertymäfunktio: Fisher ja Tippet sekä myöhemmin Gnedenko osoittivat, että tällöin F voidaan välttämättä kirjoittaa jollakin seuraavalla tavalla:

1. $F(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}$, missä $-\infty < z < \infty$; tämä on Gumbelin ääriarvojakauma.
2. $F(z) = \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\} 1_{(b,\infty)}(z)$; tämä on Fréchet'n ääriarvojakauma.
3. $F(z) = \exp\left\{-\left(-\frac{z-b}{a}\right)^{-\alpha}\right\}$ kun $z < b$ ja $F(z) = 1$, kun $z \geq b$; tämä on Weibullin ääriarvojakauma.

Tulos ei ole voimassa kaikille mahdollisille jakaumille. Todistus perustuu siihen, kuinka jakauman oikeanpuolinen häntä käyttäytyy. Yllä esitetty skaalaus lukujonojen a_n ja b_n avulla onnistuu vain jatkuvilla jakaumilla, joiden häntä ei ole liian paksu. Esimerkiksi kertymäfunktio G antaa maksimin raja-jakaumaksi Fréchetin ääriarvojakauman jos ja vain jos $G(x) < 1$ kaikilla reaaliluvuilla x, $\int_1^\infty \frac{(1-G(x))}{x} dx < \infty$ ja lisäksi on voimassa

$$\lim_{t\to\infty} \frac{\int_t^\infty \frac{(1-G(x))}{x} dx}{1 - G(t)} = \frac{1}{\alpha}.$$

Lähteessä [4] on esitetty kuinka voidaan karakterisoida se, milloin jakaumalla on maksimin asymptoottisena jakaumana joko Gumbelin jakauma tai Weibullin ääriarvojakauma.

Verrataan yllä esiteltyjä maksimin raja-arvolauseita *keskeiseen raja-arvolauseeseen*. Siinä analysoidaan riippumattomien satunnaismuuttujien skaalatun summan rajajakaumaa Olkoot X_i riippumattomia ja samoin jakautuneita satunnaismuuttujia odotusarvona $\mu = EX_1$ ja varianssina $\sigma^2 = \mathrm{Var}(X_1) = E(X_1 - \mu)^2$. Olkoon $S_n = \sum_{k=1}^n X_k$ ja Y_n normeerattu ja keskitetty summa:

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n - \beta_n}{\alpha_n};$$

keskeinen raja-arvolause kertoo, että kaikilla reaaliluvuilla pätee

$$P(Y_n \leq x) \to \Phi(x),$$

missä Φ on standartoidun normaalijakauman kertymäfunktio. Tässä oikea skaalaus on helppo löytää odotusarvon ja varianssin avulla: $\alpha_n = \sqrt{n\sigma^2}$ ja $\beta_n = n\mu$. Rajajakauma on myös yksikäsitteinen.

Maksimin asymptottinen rajajakauma on mutkikkaampi kahdellakin tavalla. Ensiksi maksimin rajajakauma ei ole yksikäsitteinen, vaan rajajakaumia on kolme. Toiseksi skaalauksessa tarvittavat lukujonot a_n ja b_n on hankala löytää. Voidaan osoittaa, että ne löydetään usein tarkastelemalla funktion $1/(1-G)$ käänteisfuntion U asymptoottista käyttäytymistä äärettömyydessä.

Esimerkki 1 *Oletetaan, että havainnot tulevat normaalijakaumasta odostusarvona* 0 *ja varianssina* 1*. Voidaan osoittaa, että tällöin pätee*

$$\lim_{n\to\infty} n\,(1 - G(a_n(z + b_n))) = e^{-x},$$

missä

$$b_n = (2\log n - \log\log n - \log(4\pi))^{\frac{1}{2}}$$

ja

$$a_n = \frac{1}{b_n}.$$

Kyseessä on siis Gumbelin ääriarvojakauma, $a = 1, b = 0$*.*

Kun halutaan soveltaa ääriarvojen teoriaa, niin pitää selvittää minkä jakauman vaikutusalueeseen havainnot kuuluvat.

Ääriarvojen tilastotiedettä

Yleistetty ääriarvojakauma

Voidaan osoittaa, että kaikki kolme erilaista ääriarvojakaumaa voidaan kirjoittaa *yleistetyn ääriarvojakauman avulla*, jonka kertymäfunktio G on muotoa

$$F(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \tag{4}$$

Parametreista oletetaan, että $-\infty < \mu < \infty$, $-\infty < \xi < \infty$ ja $\sigma > 0$. Parametreista μ on sijaintiparametri, σ on skaalaparametri ja ξ on muotoparametri. Kertymäfunktio G on määritelty niillä argumentin z arvoilla, joille pätee $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$. Fréchetin jakauma saadaan kun $\xi > 0$, Weibullin jakauma kun $\xi < 0$ ja Gumbelin jakauma rajaarvona $\xi \to 0$.

Empiirisen jakauman käyttö

Miten jakauman parametrit voidan estimoida silloin kuin saadaan havaintoja satunnaismuuttujasta X? Olkoot $X_1, X_2, \ldots$ riippumattomia ja samoin jakautuneita satunnaismuuttujia ja olkoot $x_1, x_2, \ldots$ niiden havaitut arvot. Muodostetaan havainnosta saman mittaisia blokkeja ja haetaaan blokkien maksimit $m_{n,1}, m_{n,2}, \ldots, m_{n,k}$, missä n on kunkin blokin pituus. Esimerkiksi havainnot x_j voisivat olla vedenpinnan päivittäinen korkeus, ja $m_{n,i}$ vuoden maksimikorkeus. Järjestämällä nyt havainnot $m_{n,j}$, $j = 1, \ldots, k$ suuruusjärjestykseen saadaan *empiiriseen jaukaumaan* perustuvat kvantiilit $m_{n,(j)}$, $j = 1, \ldots, k$.

Haetaan seuraavaksi yleistetyn ääriarvojakauman kvanttiilipisteet $F(z_p) = 1 - p$, $p = \frac{1}{k}, \ldots, \frac{k-1}{k}, 1$ kääntämällä yhtälö (4) ja saadaan

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{kun } \xi \neq 0 \\ \mu - \sigma \log y_p, & \text{kun } \xi = 0, \end{cases}$$

missä $y_p = -\log(1-p)$.

Piirtämällä nyt kuvaaja, missä piirreään empiirisen jakauman kvantiilipisteiden logaritmit pisteitä y_p vastaan voidaan kuvasta arvioida parametrin ξ arvo:

- Lineaarinen: $\xi = 0$. Kyseessä on siis Gumbelin jakauma.
- Konveksi: $\xi < 0$. Kyseessä on Weibullin jakauma.
- Konkaavi: $\xi > 0$. Kyseessä on Fréchetin jakauma.

Makkonen on selvittänyt graafiseen tarkasteluun liittyviä vaikeuksia [8].

Suurimman uskottavuuden estimointi

Jakauman (4) parametreja μ, σ ja ξ voi estimoidan myös suurimman uskottavuuden menetelmällä. Colesin mukaan näin saadut parametrien estimaattorit ovat säännöllisiä, kun parametri $\xi > -0.5$. Tällöin suurimman uskottavuuden estimaattorit ovat asymptoottisesti normaalisia ja niille voi johtaa asymptoottisia luottamusvälejä. Kun $-1 < \xi < -0.5$ suurimman uskottavuuden estimaattorit voidaan vielä numeerisesti löytää, mutta ne eivät enää ole sännöllisiä, joten niiden asymptoottisten jakaumaominaisuuksien selvittäminen on hankalaa. Kun $\xi < -1$, niin suurimman uskottavuuden estimaattoria on vaikea löytää. Monissa sovelluksissa kuitenkin juuri se alue, missä $\xi < 0$ on mielenkiitoinen.

Lähteessä [3] on paljon esimerkkejä, kuinka ääriarvojen tilastollista analyysiä voi käytännössä toteuttaa. Ne on toteutettu S-PLUS ohjelmalla.

Ylitystodennäköisyys

Seuraava tilanne on varsin tavallinen. Oletetaan, että meillä on riippumattomia ja samoin jakautuneita satunnaismuuttujia $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{N+n}$. Olkoon $Y_n = \max_{k \leq n} X_k$ ja $U_N = \max_{n+1 \leq k \leq N+n} X_k$. Ajatellaan, että Y_n on havaittu ja halutaan laskea todennäkoisyys sille, että U_N on suurempi kuin Y_n.

Nyt saadaan, käyttämällä hyväksi sitä, että muuttujat ovat riippumattomia ja samoin jakautuneita

$$P\left(\max_{k \leq n} X_k \in dy\right) = P(Y_n \in dy) = n\left(G\left(y\right)\right)^{n-1} G(dy) \tag{5}$$

ja muuttujien riippumattomuudesta saadaan, että

$$P(U_N \leq y | Y_n \in dy) = n\left(G\left(y\right)\right)^{N+n-1} G(dy),$$

sillä $P(U_N \leq y) = \left(G\left(y\right)\right)^N$. Saadaan siis, että

$$\begin{aligned} P(U_N \leq Y_n) &= \int_0^\infty P(U_N \leq y | Y_n \in dy) \\ &= n \int_0^\infty \left(G\left(y\right)\right)^{N+n-1} G(dy) = n \int_0^1 z^{N+n-1} dz = \frac{n}{n+N}, \end{aligned}$$

missä viimeistä edellinen yhtälö seuraa Lebesguen lemmasta [7, Lemma 1.38].

Huomaa, että esitetty tekniikka ei käytä tietoa havaitusta maksimista, ja pätee kaikille (jatkuville) jakaumille.

Ylitystodennäköisyys saadaan nyt laskettua kaavalla

$$P(U_N > Y_n) = \frac{N}{n+N}. \tag{6}$$

Esimerkki 2 *Veden korkeutta on havaittu 50 vuotta. Millä todennäköisyydellä seuraavan 20 vuoden aikana 50 vuoden maksimi ylitetään. Edellä johdetun perusteella todenäköisyydeksi saadaan*

$$p = \frac{20}{50+20} = \frac{20}{70} \sim 0.29.$$

Kaavalla voi arvoida myös muunlaisia todennäköisyyksiä. Kuinka pitkän ajan kuluessa 50 vuoden maksimia ei ylitetä todennäköisyydellä 0.9*? Saadaan yhtälö*

$$0.9 = \frac{50}{50+N},$$

mistä saadaan $N = \frac{5}{0.9} \sim 5.56$ *vuotta.*

Edellisen esimerkin arvioita voi parantaa ääriarvojakaumien avulla seuraavasti. Oletetaan, että G on jo ääriarvojakauma, esimerkiksi $G(x) = \exp[-e^{-\alpha(x-u)}]$. Oletetaan edelleen, että havaitaan n kappaleita riippumattomia satusnnaismuuttujia jakaumasta G; maksin jakauman kertymäfunktio on

$$\left(G\left(x\right)\right)^n = \exp\left[-ne^{-\alpha(x-u)}\right] = \exp\left[-e^{-\alpha(x-u-\frac{\ln(n)}{\alpha})}\right].$$

Merkitään $u_n = u + \frac{\ln(n)}{\alpha}$. Olkoon $\hat{G}_n$ *empiirinen* jakauma:

$$\hat{G}_n(x) = \frac{1}{n}\sum_{k=1}^{n} 1_{(-\infty,x)}(X_k),$$

toisin sanoen lasketaan kuinka monta havainnoista X_k on pienempiä kuin x ja jaetaan tämä havaintojen lukumäärällä n. Olkoon havaittu maksimi y_n. Selvästi $\hat{G}_n(y_n) = 1 - \frac{1}{n}$. Koska tässä jakaumassa parametri u_n on havaintojen suurin mahdollinen arvo, niin sen estimaattori on y_n. Lasketaan nyt kertymäfunktio tilanteessa missä saadaan N kappaletta uusia havaintoja:

$$\begin{aligned}
[G(x)]^N &= [(G(x))^n]^{\frac{N}{n}} = \exp\left[-\frac{N}{n}e^{-\alpha(x-y_n)}\right] \\
&= \exp\left[-e^{-\alpha(x-y_n-\frac{\ln(\frac{N}{n})}{\alpha})}\right]
\end{aligned}$$

Havaitaan, että jakauman tyyppi pysyy samana, mutta $u_N = y_n + \frac{\ln(\frac{N}{n})}{\alpha}$. Käyttämällä samoja merkintöjä kuin yllä saadaan ylitystodennäköisyydeksi

$$P(U_N > y_n) = 1 - \exp\left[-e^{-\alpha(y_n-u_N)}\right] = 1 - e^{-\frac{N}{n}}.$$

Helposti havaitaan, että $1 - e^{-\frac{N}{n}} \geq \frac{N}{N+n}$. Lisätietoa ylitystodennäköisyyksien arvioinnista ääriarvojakaumilla löytyy lähteestä [1].

Lopuksi

Artikkelissa käsiteltiin lyhyesti ääriarvojen teorian perusteita ja ääriarvoihin liittyvää tilastollista päättelyä lähinnä lähteiden [3, 4] perusteella. Perusteellinen esitys rakenteiden mekaniikan kannalta on lähteessä [2]. Lähteessä [1] on myös paljon esimerkkejä erilaisista sovelluksista.

Artikkelissa käsiteltiin yksiulotteista ääriarvojen teoriaa, mutta usein on luontevaa ajatella että rakenteiden luotettavuus riippuu esimerkiksi useasta satunnaismuuttujasta. Esimerkiksi veden pinta ylittää padon, jos sekä veden korkeus X että aallon korkeus Y ovat yhdessä tarpeeksi korkealla. Tällöin tarvitaan kaksiulotteisten satunnaismuuttujien, tai yleisemmmin moniulotteisten, satunnaismuuttujien ääriarvojen teoriaa. Lähteen [3] lisäksi näitä asioita on käsitelty lähteessä [4], joka on tärkeä yhteenveto viimeaikaisesta tilastomatemaattisesta tutkimuksesta tällä alalla. Klassisia lähteitä ääriarvojen teoriassa ovat monografit [6] ja [5].

Viitteet

[1] A. Ang, and W. Tang. *Probability concepts in Engineering Planning and Design. Volume II, Decision, Risk, and Reliability.*, Wiley, 1985.

[2] G. Augusti, A. Baratta, and F. Casciati. *Probabilistic Methods in Structural Engineering.* Chapman and Hall, 1984.

[3] S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, 2001.

[4] L. de Haan, and A. Ferreira. *Extreme Value Theory – An Introduction.* Springer 2006.

[5] B. Gnedenko. *Theory of Probability.* CRC Press, 1998.

[6] E. Gumbel. *Statistics of Extremes.* Columbia University Press, 1958.

[7] S. He, J. Wang, and J. Yan. *Semimartingale Theory and Stochastic Calculus.* CRC Press, 1992 .

[8] L. Makkonen. Bringing closure to the plotting position controversy. *Commun. Stat. Theory Methods*, **37**, 460-467, 2008.

Esko Valkeila
TKK, Matematiikan ja systeemianalyysin laitos
PL 1100, 02015 TKK
s-posti: `esko.valkeila@tkk.fi`