MATH.APP.210 Johdatus todennäköisyyslaskentaan ja tilastolliseen päättelyyn

Jussi Kangas

Sisältö (osa 5)

1. Otosjakaumien piste-estimointi
   - Otoskeskiarvo
   - Approksimaatio normaalijakaumalla
   - Otosvarianssi ja χ^2-jakauma
2. Luottamusväliestimointi
   - Luottamusvälin käsite
   - Odotusarvon luottamusväli ja t-jakauma
   - Muita luottamusvälejä

Satunnaismuuttujan otos

- tarkastellaan n-toistokoetta, jossa satunnaismuuttujalle X realisoituu jokin arvo
- jos $X_1, X_2, \ldots, X_n$ ovat satunnaismuuttujia, jotka liittyvät erillisiin toistoihin, ne muodostavat n:n kappaleen **otoksen** satunnaismuuttujasta X
- muuttujat X_i ovat riippumattomia ja noudattavat samaa jakaumaa kuin X, jolloin $E(X_i) = E(X)$ ja $\mathrm{Var}(X_i) = \mathrm{Var}(X)$
- myös realisoituneiden arvojen joukkoa $x_1, x_2, \ldots, x_n$ kutsutaan otokseksi
- jatkossa otos viittaa satunnaismuuttujan otokseen

Otossuureet

- satunnaismuuttujan X otoksesta $X_1, X_2, \ldots, X_n$ riippuva **otossuure** Θ (iso theta) on otokseen liittyvä tunnusluku, kuten otoskeskiarvo $\overline{X}$, otosvarianssi S^2 tai otoskeskihajonta S.
- otossuureet muodostetaan otosmuuttujien X_i funktioina, joten ne ovat satunnaismuuttujia
- jos ei tunneta X:n jakauman jotain parametria θ (pieni theta), voidaan sitä arvioida eli **estimoida** otossuureilla tai käyttää otossuureita parametreihin liittyvien väitteiden testaamiseen
- usein estimoidaan mm. X:n odotusarvoa ja varianssia
- parametriin θ liittyvä otossuure Θ on parametrin **estimaattori** ja sille realisoituva arvo on **(piste-)estimaatti**

Otoskeskiarvo

- jos $X_1, X_2, \ldots, X_n$ on otos satunnaismuuttujasta X ja $x_1, x_2, \ldots, x_n$ otosmuuttujien realisoituneet arvot, niin satunnaismuuttujan X **otoskeskiarvo** on satunnaismuuttuja

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

ja sen realisoitunut arvo on

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- koska $\overline{X}$ on satunnaismuuttuja, voimme tarkastella sen jakaumaa

Otoskeskiarvon jakauma

- jos satunnaismuuttujan X odotusarvo on μ ja varianssi σ^2, niin otoskeskiarvon $\overline{X}$ odotusarvo ja varianssi ovat

$$E\left(\overline{X}\right) = E\left(X\right) = \mu \qquad \text{ja} \qquad \text{Var}\left(\overline{X}\right) = \frac{\text{Var}\left(X\right)}{n} = \frac{\sigma^2}{n}$$

- otossuure Θ on **harhaton estimaattori** parametrille θ ja realisoitunut arvo on **harhaton estimaatti**, jos

$$E\left(\Theta\right) = \theta$$

- $E\left(\overline{X}\right)$ on harhaton estimaattori $E\left(X\right)$:lle

Keskiarvon keskivirhe

- otoskeskiarvon keskihajontaa kutsutaan **keskiarvon keskivirheeksi**, merkitään

$$D\left(\overline{X}\right) = \sqrt{\mathrm{Var}\left(\overline{X}\right)} = \frac{\sigma}{\sqrt{n}}$$

- otoksesta arvioitu X:n odotusarvo ja sen virhearvio esitetään usein karkeasti arvioiden muodossa $\mu \pm \sigma/\sqrt{n}$
- luotettavampi tapa on muodostaa odotusarvolle μ ns. **luottamusväli**, joka suurella todennäköisyydellä sisältää varsinaisen odotusarvon

Normaalijakautuneen satunnaismuuttujan otoskeskiarvo

- jos $X_1, X_2, \ldots, X_n$ on otos normaalijakautuneesta satunnaismuuttujasta $X \sim \mathrm{N}(\mu, \sigma^2)$, niin otoskeskiarvo $\overline{X}$ on normaalijakautunut,

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- aina ei voida olettaa, että X:llä on normaalijakauma tai että sen jakauma edes tunnetaan
- kuitenkin suurilla otoksilla otoskeskiarvon $\overline{X}$ jakauma on lähellä normaalijakaumaa (merkissä $\dot{\sim}$ piste päällä =“noudattaa likimain”) riippumatta satunnaismuuttujan X jakaumasta tai siitä onko jakauma diskreetti tai jatkuva

Normaalijakautuneen satunnaismuuttujan otoskeskiarvo

Esimerkki

PVC putkea valmistetaan asetuksilla joiden tiedetään tuottavan keskimäärin 2,565 senttiä paksua putkea hajonnan ollessa 0,0076 senttiä. Lisäksi tiedetään, että putken paksuus on normaalijakautunut. Millä todennäköisyydellä 9 kappaleen otoksessa otoskeskiarvo olisi suurempi kuin 2,56 senttiä ja pienempi kuin 2,57 senttiä?

Nyt siis $\overline{X} \sim N(2.565, 0.0076^2/9)$. Lasketaan todennäköisyys tapahtumalle $P(2.56 \leq \overline{X} \leq 2.57)$.

Matlabista komennolla

normcdf(2.57,2.565,0.0076/3)-normcdf(2.56,2.565,0.0076/3)

saadaan todennäköisyydeksi likimain 0,9516.

Keskeinen raja-arvolause

Keskeinen raja-arvolause

Olkoon $X_1, X_2, \ldots, X_n$ otos satunnaismuuttujasta X, jonka odotusarvo on μ ja varianssi σ^2. Tällöin standardoidun otoskeskiarvon

$$\overline{X}^* = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

kertymäfunktio $F(t)$ lähenee $\mathrm{N}(0,1)$-jakauman kertymäfunktiota $\Phi(t)$, kun $n \to \infty$ eli

$$F(t) = P\left(\overline{X}^* \leq t\right) \to \Phi(t), \quad \text{kun } n \to \infty.$$

Normaaliapproksimaatio

- otoskeskiarvon **normaaliapproksimaatio**: suurilla n:n arvoilla $\overline{X}^*$ on likimain standardinormaalijakautunut, $\overline{X}^* \dot{\sim} \mathrm{N}(0,1)$
- yleensä normaaliapproksimaatio antaa hyviä arvioita jo arvoilla $n \geq 30$ riippumatta X:n jakaumasta
- "lähellä" normaalijakaumaa olevilla jakaumilla voidaan soveltaa myös pienemmillä n:n arvoilla

Seuraus

Olkoon $X_1, X_2, \ldots, X_n$ otos satunnaismuuttujasta X, jonka odotusarvo on μ ja varianssi σ^2. Tällöin otoskeskiarvolle ja otoksen summalle pätevät

$$\overline{X} \dot{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad X_1 + X_2 + \cdots + X_n \dot{\sim} \mathrm{N}\left(n\mu, n\sigma^2\right)$$

Esimerkki normaaliapproksimaatiosta

Esimerkki

Hissi voi kuljettaa korkeintaan 25 henkilöä tai 2000 kg. Henkilöpaino (kg) on satunnaismuuttuja, jonka odotusarvo on $\mu = 74$ ja varianssi $\sigma^2 = 100$. Millä todennäköisyydellä satunnaisesti valitun 25 henkilön kokonaispaino ylittää 2000 kg? Jos $X =$"25 hlön kokonaispaino", niin

$$X \dot{\sim} \mathrm{N}(25 \cdot 74, 25 \cdot 100) = \mathrm{N}(1850, 2500).$$

Nyt siis kysytään, että mitä on $P(X > 2000)$. Tämä on jälleen helpointa selvittää ohjelmistolla: Matlab-komento normcdf(2000,1850,sqrt(2500),'upper') antaa todennäköisyydeksi likimain 0,0013.

Otosvarianssi

- jos $X_1, X_2, \ldots, X_n$ on otos satunnaismuuttujasta X ja $x_1, x_2, \ldots, x_n$ otosmuuttujien realisoituneet arvot, niin satunnaismuuttujan X **otosvarianssi** on satunnaismuuttuja

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

ja **otoshajonta** $S = \sqrt{S^2}$, joiden realisoituneet arvot ovat

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad \text{ja} \qquad s = \sqrt{s^2}$$

- otosvarianssi on varianssin harhaton estimaattori eli

$$E\left(S^2\right) = \mathrm{Var}(X) = \sigma^2$$

χ^2-jakauma

- normaalijakautuneiden satunnaismuuttujien muunnoksina saadaan uusia jakaumia, mm. χ^2-jakauma (χ^2 =khii toiseen).
- jatkuva satunnaismuuttuja W noudattaa **χ^2-jakaumaa vapausastein** n, $W \sim \chi^2(n)$, jos sen tiheysfunktio on

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad \text{kun } x \in \Omega = [0, \infty),$$

  missä $\Gamma(t)$ on Eulerin gammafunktio $\Gamma(t) = \int_0^\infty e^{-x} x^{t-1} dx$
- satunnaismuuttujan $W \sim \chi^2(n)$ odotusarvo ja varianssi ovat

$$E(W) = n \qquad \text{Var}(W) = 2n$$

χ^2-jakautuneiden satunnaismuuttujien summa

- jos satunnaismuuttujat $Z_i \sim \mathrm{N}(0,1)$, $i = 1,2,3,\ldots,n$ ovat riippumattomia, niin niiden neliösummalle pätee

$$W = \sum_{i=1}^{n} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$$

- jos alla esitetyt satunnaismuuttujat W_1 ja W_2 ovat riippumattomia ja jakautuneet seuraavasti

$$W_1 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n) \quad \text{ja} \quad W_2 = \sum_{i=1}^{m} U_i^2 \sim \chi^2(m),$$

niin muuttujat $Z_1, Z_2, \ldots, Z_n$, $U_1, U_2, \ldots, U_m$ ovat riippumattomia ja $W_1 + W_2 \sim \chi^2(n+m)$

Otosvarianssin jakauma

- sivun otsikosta huolimatta tutkitaan pikemminkin otosvarianssiin S^2 liittyvän otossuureen W jakaumaa

Lause

Jos $X_1, X_2, \ldots, X_n$ on otos muuttujasta $X \sim \mathrm{N}(\mu, \sigma^2)$, niin

1. $\overline{X}$ ja S^2 ovat riippumattomia
2. otossuureelle

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

pätee $W \sim \chi^2(n-1)$

Esimerkki

Lasilevyn paksuuden hajonta ei saisi ylittää arvoa $0,2$ mm. Kun $n = 40$ satunnaisesti valitun lasilevyn paksuudet mitattiin saatiin otoshajonnaksi $0,214$ mm. Paksuuden jakaumaa voidaan pitää normaalina. Mikä on todennäköisyys näinkin suurelle otoshajonnalle, jos tehtaan laitteet toimivat kuten pitää?

Tutkitaan siis kuinka suurella todennäköisyydellä otoshajonnaksi saadaan arvo 0,214 mm tai vielä suurempi arvo, jos todellisuudessa $\sigma = 0,2$ mm. Nyt

$$P(S^2 \geq 0,214^2) = P\left(\frac{(n-1)S^2}{\sigma^2} \geq \frac{39 \cdot 0,214^2}{0,2^2}\right) \approx P(W \geq 44,65)$$

χ^2-jakaumasta vapausasteella $n-1 = 39$ saadaan Matlabilla komennolla chi2cdf(44.65,39,'upper') todennäköisyydeksi likimain 0.2464.

Luottamusväliestimointi

- tilastotieteessä puhutaan usein **populaatiosta**, josta valitaan satunnaisotos ja jonka populaatiojakauma tunnetaan tai ei tunneta
- usein estimoidaan jotan populaation liittyvää parametria θ otokseen liittyvillä harhaatomilla estimaattoreilla
- tässä piste-estimoinnilla saatavaa parametrin arvon tarkkuutta ei oikein voi päätellä
- tyypillisempi estimoinnin muoto on siksi **luottamusväliestimointi**
- otoksen perusteella reaalilukuväli, jolla parametrin θ **luotetaan** olevan

Luottamusväli

- satunnaismuuttujan X parametrin θ $100(1-\alpha)\%$**:n väliestimaattori** (α=alfa, $\alpha \in [0,1]$) on satunnaisväli $[\hat{\theta}_1, \hat{\theta}_2]$, missä $\hat{\theta}_1$ ja $\hat{\theta}_2$ ovat sellaisia satunnaismuuttujan X otoksesta $X_1, X_2, \ldots, X_n$ riippuvia välin päätepisteitä, joille

$$P\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) = 1-\alpha$$

- realisoitunutta väliä $[\theta_1, \theta_2]$ sanotaan silloin parametrin θ $100(1-\alpha)\%$**:n luottamusväliksi**
- luku $1-\alpha$ on välin **luottamusaste** tai **luottamustaso**
- päätepiseet ovat välin **alempi** ja **ylempi luottamusraja**

Luottamusvälin tulkinta

- monet tekevät intuitiivisen, mutta hiukan virheellisen tulkinnan, että luottamusväli olisi reaalilukuväli, jolle parametrin arvo θ kuuluu todennäköisyydellä $1-\alpha$
- parametriin θ liittyvä väliestimaattori $[\hat{\theta}_1, \hat{\theta}_2]$ muodostuu satunnaismuuttujista $\hat{\theta}_1$ ja $\hat{\theta}_2$, joille realisoituu erilaisia arvoja otoksesta riippuen
- jos otoksia kerättäisiin hyvin monta erilaista, niin niiden perusteella lasketuista luottamusväleistä $100\,(1-\alpha)\%$ sisältää parametrin θ todellisen arvon.
- luottamusväli on siis olennaisesti satunnaisväli

Luottamusvälin valinnasta

- mitä suurempi luottamusaste vaaditaan, sitä leveämpi luottamusväli
- väliestimaattorin määritelmän ehdon $P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$ täyttäviä välejä on useita samalle luottamusasteelle $1 - \alpha$, eli määritelmä ei kerro tarkempaa perustetta välin valitaan
- useimmiten perusteena on seuraavanlainen symmetria

$$P(\theta \leq \hat{\theta}_1) = P(\theta \geq \hat{\theta}_2) = \frac{\alpha}{2}$$

- usein käytettyjä luottamusasteita ovat 90%, 95% ja 99%, jotka vastaavat **riskitasoja** $\alpha = 0.1$, $\alpha = 0.05$ ja $\alpha = 0.01$

Odotusarvon luottamusväli, kun varianssi tunnetaan

Lause

Jos $X_1, X_2, \ldots, X_n$ on otos muuttujasta, jonka **varianssi** σ^2 **tunnetaan**, niin odotusarvon $100(1-\alpha)\%$:n väliestimaattori on

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

missä $\Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$. Jos otoskeskiarvolle realisoituu arvo $\overline{x}$, niin odotusarvon $100(1-\alpha)\%$:n luottamusväli on

$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Odotusarvon luottamusväli, kun varianssi tunnetaan

Esimerkki

Tietyn muovisen liittimen paksuutta tutkittaessa otettiin $n = 75$ liittimen otos ja mitattiin paksuudet. Valmistusprosessista johtuen tiedetään paksuuden hajonnaksi $\sigma = 0,04$ mm. Otoskeskiarvo on $\overline{x} = 7,874$ mm. Etsitään paksuudelle 95 % ja 90 % luottamusvälit.

95 prosentin väliä määritettäessä $\alpha = 0,05$ eli $\alpha/2 = 0,025$.
Matlabilla komennolla norminv(1-0.025,0,1) saadaan $z_{\alpha/2} \approx 1,96$.
Luottamusväliksi kaavalla $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ saadaan (7,8649; 7,8831).

90 prosentin väliä määritettäessä $\alpha = 0,1$ eli $\alpha/2 = 0,05$.
Matlabilla komennolla norminv(1-0.05,0,1) saadaan $z_{\alpha/2} \approx 1,645$.
Luottamusväliksi kaavalla $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ saadaan (7,8664; 7,8816).

Huomioita odotusarvon luottamusvälistä

- Edellisen lauseen soveltaminen vaatii sen, että 1) jakauman varianssi σ^2 tunnetaan ja 2) otoskoko on kohtuullisen suuri (keskeisen raja-arvolauseen vaatimus)
- tavallisempi tilanne kuitenkin on, että varianssia ei tunneta, vaikka otosjakauma voidaankin olettaa normaalijakautuneeksi
- varienssin harhaton piste-estimaatti löydettiin otosvarianssin S^2 avulla, mutta koska se on satunnaismuuttuja, väliestimointiin vaikuttavien satunnaistekijöiden lisääntymisen myötä otoskeskiarvo ei ole enää normaalijakautunut
- tuntemattoman varienssin tapauksessa käytetään (Studentin) t-jakaumaa

Studentin t-jakauma

- jatkuva satunnaismuuttuja T noudattaa Studentin t-**jakaumaa vapausastein** n, $T \sim t(n)$, jos tiheysfunktio on

$$f(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad \text{kun } x \in \Omega = \mathbb{R},$$

  missä $\Gamma(t)$ on Eulerin gammafunktio $\Gamma(t) = \int_0^\infty e^{-x} x^{t-1} dx$

- tiheysfunktion kuvaaja on yksihuippuinen ja symmetrinen keskikohdan 0 suhteen
- muistuttaa normaalijakaumaa $\mathrm{N}(0,1)$ ja suurilla vapausasteilla n lähestyy ko. jakaumaa

t-jakaumaa noudattavia satunnaismuuttujia

- jos $Z \sim \mathrm{N}(0,1)$ ja $W \sim \chi^2(n)$ ovat riippumattomia, niin

$$T = \frac{Z}{\sqrt{\frac{W}{n}}} \sim t(n)$$

Lause

Olkoon $X_1, X_2, \ldots, X_n$ otos muuttujasta $X \sim \mathrm{N}(\mu, \sigma^2)$. Jos S^2 on muuttujan X otosvarianssi, niin

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Odotusarvon luottamusväli, kun varianssi on tuntematon

Lause

Jos $X_1, X_2, \ldots, X_n$ on otos muuttujasta $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, jonka **varianssi on tuntematon**, niin odotusarvon $100\,(1-\alpha)\%$:n väliestimaattori on

$$\left[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right],$$

missä $t_{\alpha/2}$ toteuttaa ehdon $P\left(T \leq t_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$, kun $T \sim t\,(n-1)$. Jos otoskeskiarvolle realisoituu arvo $\overline{x}$ ja otoshajonnalle arvo s, niin odotusarvon $100\,(1-\alpha)\%$:n luottamusväli on

$$\left[\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Esimerkki

Tietyllä seudulla aikuisten miesten pituuden tutkimiseksi valittiin satunnaisesti $n = 50$ miestä. Saatu otoskeskipituus oli $\overline{x} = 178,15$ cm ja otoshajonta oli $s = 5,92$ cm. Pituus oletetaan kyllin tarkasti normaalijakautuneeksi. Etsitään 98 % luottamusväli seudun miesten keskipituudelle.

Nyt siis $\alpha = 0,02$ eli $\alpha/2 = 0,01$. Matlabilla komennolla tinv(1-0.01,50-1) saadaan $t_{\alpha/2} \approx 2,405$. Nyt kaavalla $\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ saadaan luottamusväliksi (176,1365; 180,1635).

Huomioita odotusarvon luottamusvälistä

- kun otos hyvin suuri, on vapausasteluku $n-1$ myös suuri, joten jakauma lähellä standardinormaalijakaumaa
- tällöin otosvarianssi, joka on varianssille harhaton piste-estimaatti, tarkentuu kohti tuntematonta varianssia
- siksi kahdessa odotusarvon luottamusvälin kaavassa esiintyvät luvut $z_{\alpha/2}$ ja $t_{\alpha/2}$ lähellä toisiaan, kun otoskoko hyvin suuri
- esim. 95%:n luottamusväleille $z_{\alpha/2} \approx 1.9600$ ja otoskoolla 100 laskettu $t_{\alpha/2} \approx 1.9840$

Normaalijakauman varianssin luottamusväli

Lause

Jos $X_1, X_2, \ldots, X_n$ on otos muuttujasta $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, niin varianssin $100\,(1-\alpha)\%$:n väliestimaattori on

$$\left[\frac{(n-1)\,S^2}{w_2}, \frac{(n-1)\,S^2}{w_1}\right],$$

missä luvut w_1 ja w_2 on valittu siten, että $P\,(W < w_1) = \frac{\alpha}{2}$ ja $P\,(W > w_2) = \frac{\alpha}{2}$, kun $W \sim \chi^2\,(n-1)$. Jos varianssille realisoituu arvo s^2, niin varianssin $100\,(1-\alpha)\%$:n luottamusväli on

$$\left[\frac{(n-1)\,s^2}{w_2}, \frac{(n-1)\,s^2}{w_1}\right].$$

Esimerkki

Pakkauskone pakkaa nestettä purkkeihin. Koneen toiminnan kannalta oleellista on purkkiin pakatun nestemäärän varianssi, jonka pitäisi olla noin 100 g^2. Nestemäärä per purkki voidaan olettaa normaalijakautuneeksi. 40 purkin otoksessa saatiin otosvarianssiksi $s^2 = 151,09$ g^2. Lasketaan 90 prosentin luottamusväli varianssille.

Nyt siis $\alpha = 0,1$ ja $\alpha/2 = 0.05$. Matlabilla saadaan χ^2-jakaumasta vapausasteella $40-1$ komennoilla chi2inv(0.05,39) ja chi2inv(0.95,39) luvuiksi w_1 ja w_2 likimain arvot 25,6954 ja 54,5722. Täten luottamusväliksi saadaan

$$\left[\frac{(n-1)s^2}{w_2}, \frac{(n-1)s^2}{w_1}\right] \approx \left[\frac{39 \cdot 151,09}{54,5722}, \frac{39 \cdot 151,09}{25,6954}\right]$$

eli likimain (107,98; 229,32). Pakkauskone voisi siis olla säätämisen tarpeessa.

Suhteellinen osuus

- oletetaan, että $X \sim \mathrm{Bin}(n, p)$, missä p on onnistumisen todennäköisyys
- todennäköisyyden frekvenssitulkinnassa p on myös onnistumisten suhteellinen frekvenssi eli niiden **suhteellinen osuus** kaikista toistoista
- tulkitaan X jakaumaa $\mathrm{Ber}(p) = \mathrm{Bin}(1, p)$ noudattavien satunnaismuuttujien summana $X = Y_1 + Y_2 + \cdots + Y_n$
- tällöin satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n$ muodostavat otoksen satunnaismuuttujasta $Y \sim \mathrm{Bin}(1, p)$
- monesti satunnaismuuttujan Y parametri p on tuntematon, joten sitä on estimoitava

Suhteellisen osuuden estimointi

- edellisen sivun satunnaismuuttujan Y otoskeskiarvo on

$$\hat{P} \overset{\text{merk.}}{=} \overline{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{X}{n}$$

- binomijakautunut X kuvaa onnistumisten frekvenssiä n-toistokokeessa, joten $\hat{P} = \frac{X}{n}$ on suhteellinen osuus
- satunnaismuuttuja $\hat{P}$ on harhaton estimaattori parametrille p, sillä otoskeskiarvon odotusarvo on

$$E(\hat{P}) = E(\overline{Y}) = E(Y) \overset{Y \sim \text{Bin}(1,p)}{=} 1 \cdot p = p$$

- seuraavan sivun luottamusväli on parametrille p hyvä estimointi, jos otoskoko on riittävän suuri

Suhteellisen osuuden luottamusväli

Lause

Oletetaan, että $X \sim \mathrm{Bin}(n, p)$. Parametrin p $100(1-\alpha)\%$:n väliestimaattori on

$$\left[\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right],$$

missä $\Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$. Jos suhteelliselle osuudelle $\hat{P}$ realisoituu arvo $\hat{p}$, niin parametrin p, niin $100(1-\alpha)\%$:n luottamusväli on

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Esimerkki

Vuoden 2024 helmikuun alussa Helsingin Sanomat teetti gallupin koskien ihmisten äänestyskäyttäytymistä presidentinvaalien tulevalla toisella kierroksella. Gallupissa haasteltiin 1154 henkilöä ja tulosten mukaan noin 57 prosenttia vastaajista äänestäisi tulevalla vaalien toisella kierroksella Alexander Stubbia. Lasketaan 95 prosentin luottamusväli Stubbin äänestäjien osuudelle äänioikeutettujen joukossa.

Nyt siis $\alpha/2 = 0{,}025$ ja arvoksi $z_{\alpha/2}$ saadaan likimain 1,96. Luottamusväliksi saadaan kaavalla

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

likimain (0,5414; 0,5986). (Huom. Lopullinen Stubbin ääniosuus 51,6 % ei siis mahtunut tälle luottamusvälille.)