5A Luottamusvälien määrittäminen

Normaalijakauman kertymäfunktion voi laskea R:ssä funktiolla `pnorm`, ja sen käänteisfunktion funktiolla `qnorm`, esimerkiksi `pnorm(1.96)` ≈ 0.975, ja vastaavasti `qnorm(0.975)` ≈ 1.96.

Matlabissa ja Octavessa vastaavat komennot ovat `normcdf` ja `norminv`.

Jos tietokonetta tai normaalijakauman osaavaa laskinta ei ole käytettävissä, kertymäfunktion arvoja voi etsiä taulukoista (ks. kurssisivu, kohta Materiaalit).

Tuntitehtävät

5A1 (Limuautomaatti) Limuautomaatti laskee mukiin juomaa määrän (ml), joka noudattaa likimain normaalijakaumaa odotusarvona μ ja keskihajontana $\sigma = 3$. Automaattia testatessa mitattiin mukeihin valutetuiksi juomamääriksi (ml): 304, 298, 301, 302, 301, 300, 305, 300, 306.

(a) Määritä 95% luottamustason väliestimaatti parametrille μ.

(b) Määritä 99% luottamustason väliestimaatti parametrille μ.

(c) Kun suoritetaan koe, jossa otetaan 9 mukia juomaa ja lasketaan luottamusväli a-kohdan mukaisesti, mikä on todennäköisyys, että luottamusväli (i) sisältää arvon μ, (ii) on kokonaan μ:n alapuolella, (iii) on kokonaan μ:n yläpuolella?

(d) Mitä enemmän dataa on saatavilla, sen kapeampi luottamusväli saadaan. Kuinka monta mittausta vaadittaisiin, että 95% luottamustason väliestimaatti saataisiin alle 1 ml levyiseksi (0.5 ml kumpaankin suuntaan)? Entä 0.1 ml levyiseksi?

Ratkaisu.

(a) Havaitusta datajoukosta $\vec{x}$ lasketaan keskiarvoksi $m(\vec{x}) \approx 301.89$. Kysytty väliestimaatti on

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}},$$

missä $n = 9$, $\sigma = 3$ ja $z > 0$ on luku, jolle normitettua normaalijakaumaa noudattava Z toteuttaa $\mathrm{P}(|Z| \leq z) = 0.95$. Toisin sanoen häntätodennäköisyyksien (vasen ja oikea yhteensä) tulee olla $\mathrm{P}(|Z| > z) = 0.05$. Koska standardinormaalijakauma on origon suhteen symmetrinen, ovat vasen ja oikea häntä yhtäsuuret, $\mathrm{P}(Z < -z) = \mathrm{P}(Z > z) = 0.05/2 = 0.025$.

Etsitään siis lukua z, jolla kertymäfunktion arvo on $F_Z(z) = \Phi(z) = 1 - 0.025 = 0.975$. Taulukoista, tai R-komennolla `qnorm(0.975)` saamme tuloksen $z \approx 1.96$.

Kysytty 95% luottamustason väliestimaatti on siis

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} \;=\; 301.89 \pm 1.96\frac{3}{\sqrt{9}} \;=\; 301.89 \pm 1.96.$$

(b) Kuten a-kohdassa, mutta jotta keskialueen todennäköisyys on 0.99, on (standardinormaalijakauman) vasemman ja oikean hännän oltava yhteensä 0.01, eli kummankin hännän 0.005. Etsitään siis sellainen luku $z > 0$, että $\Phi(z) = 0.995$. Taulukoista tai R-komennolla `qnorm(0.995)` saamme luvun $z \approx 2.58$, joten 99% luottamustason väliestimaatti on

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} \;=\; 301.89 \pm 2.58\frac{3}{\sqrt{9}} \;=\; 301.89 \pm 2.58.$$

(c) (i) 95% (ii) 2.5% (iii) 2.5%

(d) Jos 95% luottamusväli lasketaan n alkion datajoukolle $\vec{x}$, niin (a)-kohdan perusteella luottamusväliksi saadaan

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} \;=\; m(\vec{x}) \pm 1.96\frac{3}{\sqrt{n}}$$

Tämän luottamusvälin leveys on

$$2 \times 1.96\frac{3}{\sqrt{n}}$$

Luottamusvälin leveys saadaan alle 1 (ml) levyiseksi valitsemalla n siten, että

$$2 \times 1.96\frac{3}{\sqrt{n}} < 1$$

eli

$$n > (2 \times 1.96 \times 3)^2 \;=\; 138.3.$$

Limuautomaatista tulee siis valuttaa vähintään 139 mukillista.

Vastaavasti 0.1 ml levyinen luottamusväli saadaan noin 13900 mukillisella.

5A2 (Mielipidemittaus) Helsingin Sanomien heinäkuussa 2016 raportoiman kyselytutkimuksen mukaan 89 prosenttia suomalaisista oli sitä mieltä, että presidentti Niinistö on suoriutunut tehtävästään erittäin tai melko hyvin. Kysely toteuttiin haastattelemalla puhelimitse 1002 suomalaista ikähaarukassa 15–79 vuotta ja virhemarginaalin kerrottiin olevan noin 3 prosenttiyksikköä suuntaansa. Oletetaan, että virhemarginaali on laskettu käyttämällä binaarimallin konservatiivista väliestimaattoria (luentomoniste, kaava (8.5)).

(a) Päättele annettujen tietojen perustella, mitä luottamustasoa kyselytutkimusten virhemarginaalin raportoinnissa käytettiin.

(b) Kuinka monta suomalaista olisi pitänyt haastatella, jos virhemarginaaliksi olisi samalla luottamustasolla haluttu noin 1 prosenttiyksikkö suuntaansa?

Ratkaisu.

(a) Binaarimallin likimääräinen luottamusväli voidaan kirjoittaa muodossa

$$\hat{p} \pm z \times \frac{0.5}{\sqrt{n}},$$

missä $\hat{p}$ on ykkösten osuus havaitussa datajoukossa ja luku z toteuttaa

$$\Phi(z) - \Phi(-z) \;=\; \mathrm{P}(|Z| \leq z) \;=\; \beta,$$

missä β on käytetty luottamustaso.

HS:n kyselytutkimuksen tulokseksi raportoitiin luottamusväliksi 0.89 ± 0.03 otoskoolla $n = 1002$. Näin ollen parametrin p (Niinistöön tyytyväisten suomalaisten osuus) luottamusväli on

$$0.89 \pm 0.03 \;=\; 0.89 \pm z \times \frac{0.5}{\sqrt{1002}}.$$

Ratkaisemalla yhtälö

$$0.03 \;=\; z \times \frac{0.5}{\sqrt{1002}},$$

saadaan

$$z \;=\; 0.03 \times \frac{\sqrt{1002}}{0.5} \;=\; 1.90.$$

Tämä vastaa luottamustasoa

$$\beta = \mathrm{P}(|Z| \leq 1.90) \;=\; \Phi(1.90) - \Phi(-1.90) \;\approx\; 94.3\%$$

Tästä voidaan arvata, että HS-gallupissa on käytetty yleistä 95% luottamustasoa.

Voidaan vielä tarkistaa, että 95% luottamustasolla $z = 1.96$ ja tällöin likimääräinen luottamusväli on

$$0.89 \pm 1.96 \frac{0.5}{\sqrt{1002}} \;=\; 0.89 \pm 0.031.$$

(b) Jos likimääräisen 95% luottamustason väliestimaatin leveydeksi halutaan 1% suuntansa, niin tällöin

$$0.89 \pm 0.01 \;=\; 0.89 \pm 1.96 \times \frac{0.5}{\sqrt{n}}.$$

Ratkaisemalla

$$0.01 \;=\; 1.96 \times \frac{0.5}{\sqrt{n}}$$

saadaan

$$n \;=\; \left(1.96 \times \frac{0.5}{0.01}\right)^2 \;=\; 9604.$$

Luottamustasolla 95% tarvitsisi "virhemarginaali 1 prosenttiyksikköä suuntaansa" -tarkkuuden saamiseksi haastatella noin 9600 suomalaista.

Kotitehtävät

5A3 (Monen puolueen kyselytutkimus) Eräästä suuresta populaatiosta otettiin $n = 100$ henkilön kokoinen satunnaisotos, jolta kysyttiin mitä neljästä puolueesta A,B,C,D vastaajat kannattavat. Kyseisten puolueiden kannattajia oli otoksessa 80, 18, 2 ja 0 henkilöä.

(a) Käsittele kunkin puolueen X kohdalla erikseen kysymystä "mikä osuus populaatiosta kannattaa puoluetta X" binaarisena kysymyksenä (ks. luentomonisteen luku 8.3) ja laske 95% luottamusväli kyseisen puoleen kannatukselle populaatiossa. Ilmoita välit alku- ja loppupisteen avulla kolmella desimaalilla esim. muodossa [0.500, 0.600].

(b) Toista edellisen kohdan laskut käyttäen konservatiivista väliestimaattoria.

(c) Ovatko edellisissä kahdessa kohdassa lasketut välit mielekkäitä? Jos eivät, millä tavalla eivät ja mitä arvelet syyksi? Pohdi mitä asialle voisi tehdä. Opastus: Pieni pohtiminen riittää. Tällaisiin tilanteisiin on olemassa kirjallisuudessa ratkaisuja, mutta ne ovat hiukan mutkikkaampia kuin tällä kurssilla esitetyt kaavat.

(d) Onko mahdollista, että puolueen kannatus populaatiossa on nollaa suurempi, mutta otoksessa sen osuus on nolla? (Järkeile, älä laske.)

(e) Onko mahdollista, että puolueen kannatus populaatiossa on tasan nolla, mutta otoksessa sen osuus on nollaa suurempi? (Järkeile, älä laske.)

Arviointiohje.
a, b: 1/2 pistettä per kohta.
c, d, e: 1/3 pistettä per kohta.
Summa pyöristetään ylöspäin. C-kohdassa riittää jonkinlainen järkeenkäypä tarkastelu, esim. huomio luottamusvälin ulottumisesta negatiiviselle puolelle.

Ratkaisu.

(a) Puolueen A kannatuksen piste-estimaatti on $\hat{p} = 80/100 = 0.800$. Luottamustasoa 95% varten käytämme arvoa $z = 1.96$, jolloin luentomonisteen kaavassa (8.3) virhemarginaaliksi saadaan $1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n} \approx 0.078$. A-puolueen luottamusväliksi saadaan

$$[0.800 - 0.078,\ 0.800 + 0.078] = [0.722, 0.878].$$

Puolueelle B saadaan virhemarginaali 0.075 ja väli $[0.105, 0.255]$.

Puolueelle C saadaan virhemarginaali 0.027 ja väli $[-0.007, 0.047]$.

Puolueelle D saadaan virhemarginaali 0.000 ja väli $[0.000, 0.000]$.

(b) Piste-estimaatit $\hat{p}$ ovat kuten edellä, mutta kaikille puolueille käytetään samaa "konservatiivista" virhemarginaalia $z \cdot 0.5/\sqrt{n} = 1.96 \cdot 0.5/\sqrt{100} \approx 0.098$. Luottamusvälit ovat:

- A: $[0.702, 0.898]$

- B: $[0.082, 0.278]$
- C: $[-0.078, 0.118]$
- D: $[-0.098, 0.098]$

Kaikilla puolueilla virhemarginaali leveni edelliseen kohtaan verrattuna (koska käytettiin konservatiivista leveyttä).

(c) Puolueiden A ja B luottamusvälit vaikuttavat järkeviltä, mutta puolueiden C ja D vaikuttavat omituisilta. Puolueen kannattajien osuus populaatiossa ei varmasti ole negatiivinen, joten tuntuu turhalta, että luottamusväli ulottuu nollan alapuolelle. Pääasiallinen syy tälle omituisuudelle on, että käytetty kaava perustuu binomijakauman approksimoimiseen normaalijakaumalla; lähellä binomijakauman päätepisteitä tämä approksimaatio on hyvin huono.

Ainakin yksi asia on helppo tehdä. Jos esim. väli $[-0.007, 0.047]$ sisältää C:n oikean kannatusosuuden, niin myös lyhyempi väli $[0, 0.047]$ sisältää sen. Ilmeinen ajatus olisi ainakin katkaista väleistä pois negatiivinen osuus. Muita ongelmia: Puolueelle D saatiin c-kohdassa luottamusväliksi yksi piste, mikä vaikuttaa intuitiivisesti liiankin kapealta (ja onkin). Toisaalta d-kohdan konservatiivinen luottamusväli on hyvinkin konservatiivinen (löysä, leveä väli) ja antaa kovin epämääräisen kuvan D:n kannatuksesta.

Valitettavasti ongelmat eivät lopu tähän. Koska binomijakauma poikkeaa lähellä päätepisteitään huomattavasti normaalijakaumasta, niin myös luottamusvälien yläpäät ovat melkoisesti pielessä. Tässä kohdaa tyydymme toteamaan, että otososuuden ollessa lähellä nollaa tai ykköstä binaarimallin luottamusvälit kannattaa laskea jollakin tarkemmalla kaavalla, joita löytyy kirjallisuudesta, ks. esim. Brown, Cai ja DasGupta (2001): Interval estimation for a binomial proportion, *Statistical Science* 16:101–133. Näiden kaavojen osaamista ei tällä kurssilla vaadita, mutta on hyvä mieltää niiden olemassaolo. — Toinen, ehkä helpompi vaihtoehto on käyttää luottamusvälien sijasta bayesläisiä uskottavuusvälejä (seuraavat luennot).

(d) Kyllä. Näin voi helpostikin käydä, jos otos on pieni ja/tai puolueen kannatus väestössä on pieni.

(e) Ei. Nollasta kannattajasta (väestössä) ei voi saada otokseen yhtään kannattajaa.

Kohdista d–e näemme arkijärjellä seuraavaa: Jos otoksessa on $\hat{p} = 0$, tästä ei seuraa (ainakaan varmasti) että populaatiossa olisi $p = 0$. Mutta kylläkin jos otoksessa on $\hat{p} > 0$, niin *varmasti* myös populaatiossa on $p > 0$.

5A4 (Alkeishiukkaset) Eräässä fysikaalisessa kokeessa havaitaan hiukkasten hajoamisia satunnaisin välein. Väliajat oletetaan riippumattomiksi ja eksponenttijakautuneiksi tuntemattomalla odotusarvolla μ, jolloin kyseisen jakauman taajuusparametri on $\lambda = 1/\mu$. Mitattiin 30 peräkkäistä hajoamisten väliaikaa. Havaittujen väliaikojen keskiarvo oli 13.15 sekuntia ja keskihajonta 12.18 sekuntia.

(a) Käyttäen yleistä odotusarvon luottamusväliä (monisteen luku 8.2 / luento 4B), määritä 90% luottamusväli odotusarvoparametrille μ. Ilmoita tulos kahdella desimaalilla. Huom. vaadittu luottamustaso.

(b) Yksittäiset hiukkashajoamiset eivät olleet normaalijakautuneita. Miksi tässä käytetty menetelmä on kuitenkin likimain oikea?

(c) Voisiko samaa menetelmää käyttää luotettavasti, jos havaintoja (hajoamisten väliaikoja) olisi vain $n = 3$ kappaletta? Miksi / miksi ei?

Arviointiohje.
a-kohta 1p, muut 0.5p/kohta.

Ratkaisu.

(a) 90% luottamustasoa varten tarvitsemme sellaisen luvun $z > 0$, että (standardinormaalijakauman) vasen ja oikea häntä ovat kumpikin 5%, toisin sanoen $\Phi(z) = 0.95$. Taulukoista tai R-komennolla `qnorm(0.95)` saamme luvun $z = 1.64$.

Lauseen 8.2 avulla luottamusväliksi saadaan

$$m(\vec{x}) \pm z \cdot \frac{\mathrm{sd}(\vec{x})}{\sqrt{n}} = 13.15 \pm 1.64 \cdot \frac{12.18}{\sqrt{30}} \approx 13.15 \pm 3.66,$$

tai välimuodossa ilmoitettuna $[9.49, 16.81]$.

(b) Havainnot ovat 30 riippumatonta satunnaislukua samasta (eksponentti)jakaumasta. Keskeisen raja-arvolauseen nojalla monen riippumattoman luvun summa (ja siten myös keskiarvo) on likimain normaalijakautunut.

(c) Kolmen eksponenttijakautuneen luvun summa (tai keskiarvo) ei ole kovinkaan lähellä normaalijakautunutta, joten menetelmä ei luultavasti sovellu.

Kun jakauman muoto on tässä tunnettu tai oletettu (eksponenttijakauma), voisi kirjallisuudesta yrittää etsiä juuri tähän tilanteeseen sopivia menetelmiä. Kolmen eksponenttijakautuneen muuttujan summan jakauma nimittäin tiedetään (gammajakauma). — Toinen ratkaisu on käyttää Bayes-päättelyä, joka soveltuu hyvin myös pienille otoksille.