MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

4B Tilastolliset luottamusvälit

Emilia Blåsten

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2021–2022
Periodi II

Esim. Kahviautomaatti

Haluttiin selvittää, kuinka paljon kahviautomaatti keskimäärin laskee kuppiin kahvia. Toimintaa testattiin valuttamalla automaatista 25 kupillista ja mittaamalla kahvin määrät kupeissa.

Mittauksessa havaittiin arvot (senttilitroina):
$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43,$
$10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.50, 9.38, 9.98)$

Mittausdatan keskiarvo on $m(\vec{x}) = 10.0284$.

Kysymys: Onko kahviautomaatin valuttamien kahvimäärien "todellinen keskiarvo" μ lähellä lukua 10.0284?

"Todellisella keskiarvolla" tarkoitamme tässä sen jakauman odotusarvoa, josta kahvimäärät määräytyvät satunnaisesti kullakin valutuskerralla.

"Mittausdatassa" ei (tässä tapauksessa) ajatella olevan epävarmuutta, olemme mitanneet kahvimäärät tarkasti. Epävarmuus koskee sitä, mitä kahviautomaatti yleensä/jatkossa tekee. (Yleistäminen datasta populaatioon)

Sisältö

Datalähteen stokastinen malli

Havaittu data

Datalähteestä on havaittu arvot $x_1, \ldots, x_n$. Halutaan päätellä (=arvata) tutkittavan suureen (tuntematon) jakauma $f(x)$.

Stokastinen malli

Tilastokokeen mahdollista tulosta mallinnetaan satunnaismuuttujilla $X_1, \ldots, X_n$, jotka ovat toisistaan riippumattomat ja noudattavat (tuntematonta tai oletettua) jakaumaa $f(x)$.

Stokastinen malli (generoiva malli, generoiva jakauma) kuvaa datalähteen toimintaa yleensä (millaisia lukuja se voi tuottaa ja millä tn:llä).

Datalähteen stokastinen malli

Stokastinen malli on matemaattinen yksinkertaistus datalähteen toiminnasta.

- Mallissa voi olla parametreja. Jos parametrien arvot kiinnitetään, saadaan tietty jakauma, esim. $\mathrm{Bin}(10, 0.5)$. Osa parametreista voi olla tunnettuja ja osa tuntemattomia.
- Hyvä malli kuvaa riittävän tarkasti sitä, millaisia lukuja datalähteestä tulee ja millaisin todennäköisyyksin.
- Toisaalta hyvä malli on riittävän yksinkertainen, jotta sillä voidaan riittävän helposti laskea.
- Yleensä oletetaan, että samasta lähteestä voidaan ottaa paljon lukuja ja ne määräytyvät toisistaan riippumattomasti. (Jos näin ei ole, tarvitaan monimutkaisempi malli.)

Stokastinen malli, esimerkkejä

Esimerkki (Kahviautomaatti)

Kahviautomaatista saadun kahvimäärän oletetaan määräytyvän eräästä tuntemattomasta jakaumasta, jonka odotusarvo on μ. Jakauma pyrkii kuvaamaan kahvin annostelun fysikaalista prosessia, johon vaikuttaa koneen konstruktio, säädöt ja joka kerta erikseen toteutuvat yksityiskohdat, joita emme osaa tarkasti ennustaa. Siksi kuvaamme niitä satunnaismuuttujalla.

Esimerkki (Otanta populaatiosta)

Suomalaisia on N kpl ja heistä täsmälleen K kpl eli eräs osuus $p = K/N$ kannattaa ydinvoiman lisärakentamista.
Käytännön syistä poimimme satunnaisen suomalaisen, jolloin hänen ydinvoiman kannatustaan kuvaa satunnaismuuttuja $X_1 \sim \text{Ber}(p)$, missä parametri p on vakio, jonka arvoa emme tunne. Voimme myös poimia lisää satunnaisia suomalaisia $X_2, X_3, \ldots$.

Pienet ja isot kirjaimet (eräs konventio)

Datajoukko $\vec{x} = (x_1, \ldots, x_n)$

- Koostuu mittaamalla havaituista luvuista
- Määrittämiseen ei tarvita mitään matemaattista mallia
- Esim. $(x_1, x_2, x_3) = (10.17, 11.23, 9.59)$, kahviautomaatin kolme ensimmäistä mittausta

Stokastinen malli $\vec{X} = (X_1, \ldots, X_n)$

- Koostuu satunnaismuuttujista ja perustuu valittuun matemaattiseen malliin, jolla pyritään ennakoimaan datalähteen tuottamia arvoja
- Määrittämiseen ei tarvita lainkaan mittausdataa
- Esim. että (X_1, X_2, X_3) ovat toisistaan riippumattomia normaalijakautuneita satunnaismuuttujia odotusarvolla 10 ja keskihajonnalla 3. Tämä on eräs datalähde.

Datajoukon ja stokastisen mallin tunnusluvut

Deskriptiivisestä tilastotieteestä: tunnusluku on datasta $\vec{x} = (x_1, \ldots, x_n)$ jollakin säännöllä $g : \mathbb{R}^n \to \mathbb{R}$ laskettava yksittäinen luku.

Esim (Eräitä tunnuslukuja)

- Keskiarvo $m(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Varianssi $\mathrm{var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))^2$
- Keskihajonta $\mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})}$

Jos sääntöä (funktiota) sovelletaan stokastisen mallin mukaiseen satunnaismuuttujaan $\vec{X} = (X_1, \ldots, X_n)$, niin tulos

$$g(X_1, \ldots, X_n)$$

on satunnaismuuttuja. Stokastiikan menetelmillä voimme ymmärtää sen jakauman, ts. millaisia arvoja tunnusluku voi saada. Nyt tarvitaan aiempia tietojamme satunnaismuuttujan muunnoksen jakaumasta.

Millainen on (stok. mallin) keskiarvon jakauma?

Hypoteettista jakaumaa $f(x)$ odotusarvona μ ja keskihajontana σ noudattavan datalähteen stokastisen mallin $\vec{X} = (X_1, \ldots, X_n)$ keskiarvo $m(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ on satunnaisluku:

$$\mathbb{E}[m(\vec{X})] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$\mathrm{SD}[m(\vec{X})] = \mathrm{SD}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{SD}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sigma\sqrt{n} = \frac{\sigma}{\sqrt{n}}.$$

Estimaatin virhe ja virheen jakauma

Stok. malli: $X_1, \ldots, X_n$ riippumattomia satunnaislukuja, joilla odotusarvo μ ja keskihajonta σ.

Käytetään havaittua suuretta $m(\vec{X})$ estimaattorina tuntemattomalle suureelle μ. Mikä on estimaatin virhe ja miten virhe on jakautunut?

Tiedämme ainakin että $\mathbb{E}[m(\vec{X})] = \mu$ ja $\mathrm{SD}[m(\vec{X})] = \frac{\sigma}{\sqrt{n}}$.

Lineaarisuuden nojalla virheen $m(\vec{X}) - \mu$ odotusarvo on nolla ja keskihajonta sama kun yllä.

Mennään vielä askel pidemmälle. Jaetaan virhe omalla keskihajonnallaan, saadaan normitettu virhe

$$\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}.$$

Lineaarisuuden nojalla sen odotusarvo on 0 ja keskihajonta 1.

Miksi tämä on hyödyllistä? Koska voimme ehkä laskea todennäköisyyksiä sille, että normitettu virhe on pieni tai suuri.

Entä muiden tunnuslukujen jakaumat?

Datasta voi laskea monenlaisia tunnuslukuja, muitakin kuin keskiarvon. Myös muiden tunnuslukujen jakauma voi olla tarpeen ymmärtää. Esimerkiksi, jos jokainen havainto X_i tulee samasta jakaumasta f, niin mikä on ...

- maksimin $\max\{X_1, \ldots, X_n\}$ jakauma?
- varianssin $\text{sd}(\{X_1, \ldots, X_n\})$ jakauma?
- mediaanin $\text{med}(\{X_1, \ldots, X_n\})$ jakauma?

Esim. havaitun datan keskihajona $\text{sd}(\vec{x})$ on jokin luku, joka on ehkä lähellä datalähteen keskihajontaa $\text{SD}(X_i)$, eli estimoi sitä. Mutta kuinka lähellä? $\rightarrow$ Tarvitaan stokastiikan menetelmiä.

Tällä luennolla kuitenkin keskitymme yhteen tunnuslukuun (keskiarvoon).

Sisältö

Esim. Kahviautomaatti

Kahviautomaatin on tarkoitus laskea jokaiseen kuppiin keskimäärin 10.0 cl kahvia. Kahviautomaatin toimintaa testattiin valuttamalla automaatista 25 kupillista ja mittamalla kahvin määrät kupeissa.

Mittauksessa havaittiin arvot (cl):
$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Mittausdatan keskiarvo on $m(\vec{x}) = 10.03$. Määritä havaitun datan pohjalta luottamusväli todelliselle μ:n arvolle.

(Emme tosin ole vielä määritelleet mitä tarkoittaa "luottamusväli".)

Yleiskäsitteitä: Piste-estimaatti ja väliestimaatti

Olkoon tuntematon parametri θ.

Parametrin piste-estimaatti on jokin luku $\hat{\theta}$, joka toivottavasti on lähellä oikeaa arvoa: $\hat{\theta} \approx \theta$.

Parametrin väliestimaatti on jokin väli $[a, b]$, joka toivottavasti sisältää oikean arvon: $[a, b] \ni \theta$.

"Toivottavasti" ja "lähellä" pitää määritellä tarkemmin matemaattisesti (eri tilanteissa hiukan eri määritelmiä).

- Tällä luennolla eräs väliestimaatti: luottamusväli
- Ensi viikolla toisenlaisia, ns. bayesiläisiä väliestimaatteja

Normaalimallin odotusarvoparametrin piste-estimaatti

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Datalähteen stokastinen malli: $X_1, \ldots, X_{25}$ riippumattomia ja normaalijakautuneita odotusarvona μ ja keskihajontana $\sigma = 0.5$

Tehtävä: Estimoi normaalimallin parametri μ

Uskottavuusfunktio

$$f(x_1, \ldots, x_n \,|\, \mu, \sigma) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$\Longrightarrow$ Parametrin μ suurimman uskottavuuden estimaatti on

$$m(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; 10.03$$

Tämä on eräs piste-estimaatti, mutta kuinka tarkka?
Mikä on tn, että datalähteen tuottamia arvoja mallintavasta satunnaisvektorista laskettu keskiarvo $m(\vec{X})$ on "lähellä" parametria μ?

Normaalimallin keskiarvo

Normaalimalli: $X_1, \ldots, X_n$ riippumattomia ja normaalijakautuneita satunnaislukuja odotusarvona μ ja keskihajontana σ

Normitetun virheen

$$\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}$$

odotusarvo on 0 ja keskihajonta 1.

Koska

- riippumattomien normaalijakautuneiden summa on normaalijakautunut,
- normaalijakautuneen satunnaismuuttujan siirretty ja skaalattu versio on normaalijakautunut,

noudattaa normitettu virhe standardinormaalijakaumaa $N(0, 1)$.

Normaalimallin väliestimaatti

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Datalähteen stokastinen malli: $X_1, \ldots, X_{25}$ riippumattomia ja normaalijakautuneita odotusarvona μ ja keskihajontana $\sigma = 0.5$

$$\mathbb{P}(|m(\vec{X}) - \mu| \leq 0.2) = \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{0.2}{0.5/\sqrt{25}}\right) = \mathbb{P}(|Z| \leq 2) \approx 95\%.$$

Melko suurella todennäköisyydellä (tn $= 95\%$) siis pätee

$$\mu \in [m(\vec{X}) - 0.2, m(\vec{X}) + 0.2]$$

Havaitusta datajoukosta $\vec{x}$ laskettu

- parametrin μ piste-estimaatti on $m(\vec{x}) = 10.03$
- parametrin μ väliestimaatti on $m(\vec{x}) \pm 0.2 = [9.83, 10.23]$

Voidaanko päätellä, että väli $[9.83, 10.23]$ peittää μ:n 95% tn:llä?
Ei voida.

Väliestimaatin tulkinta

$\vec{x} = (10.17, 11.23, 9.59, 8.94, 10.14, 9.66, 10.22, 9.59, 11.11, 9.94, 9.76, 9.92, 10.43, 10.05, 9.19, 10, 10.38, 10.02, 10.37, 9.93, 9.97, 10.24, 10.5, 9.38, 9.98)$

Lukuväli

$$m(\vec{x}) \pm 0.2 \;=\; [9.83, 10.23]$$

on parametrin μ väliestimaatti luottamustasolla 95%

Normaalimallin väliestimaatti $m(\vec{X}) \pm 0.2$ auttaa ennakoimaan, millä tn datalähteen tuottamista arvoista laskettava väliestimaatti peittää μ:n:

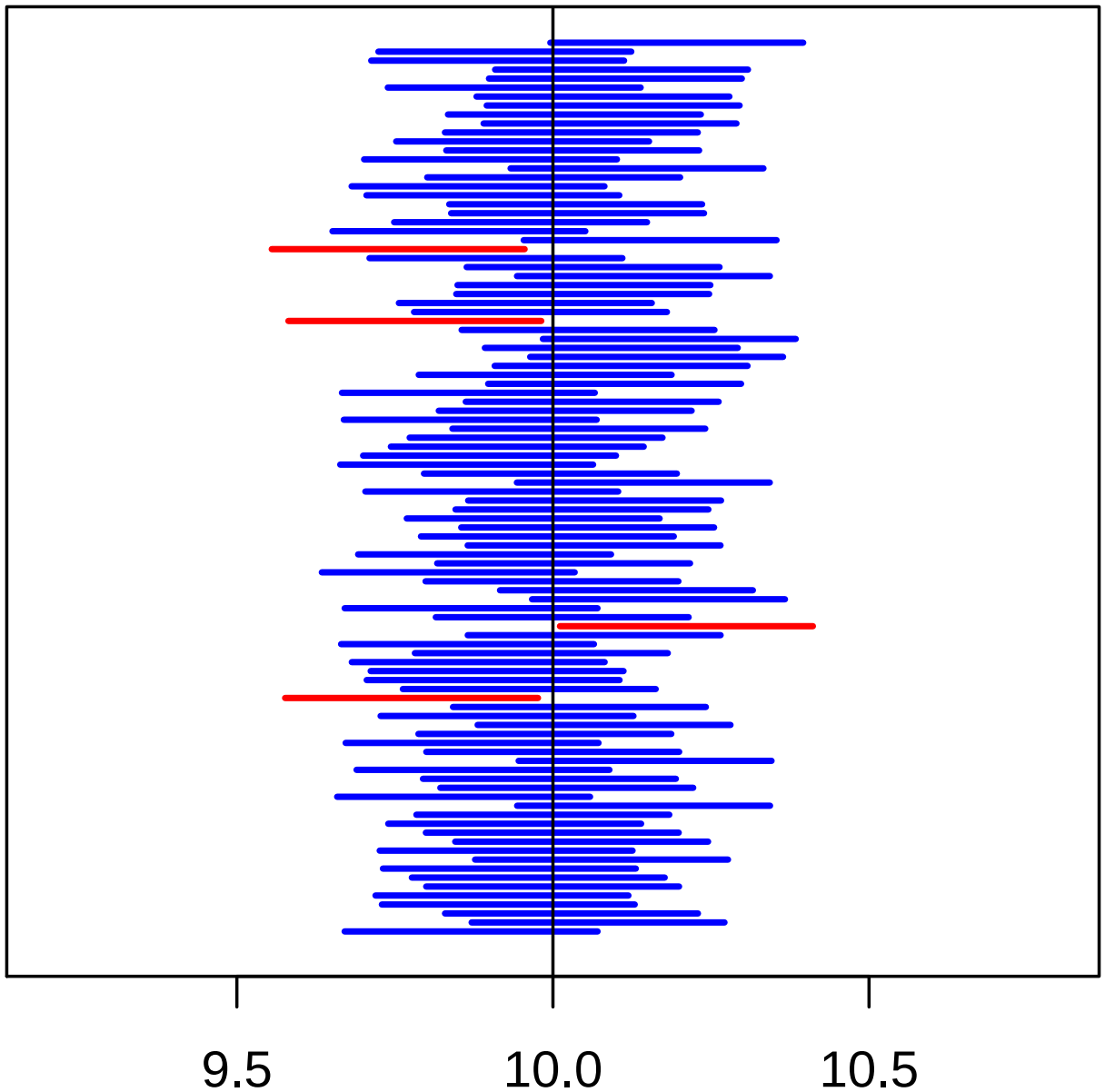
$$\mathbb{P}(\mu \in [m(\vec{X}) - 0.2, m(\vec{X}) + 0.2]) \;=\; 95\%.$$

Jo havaitusta datasta lasketun väliestimaatin $[9.83, 10.23]$ todennäköisyyksistä ei normaalimalli kerro mitään.

Henkilö, joka laskee paljon estimaatteja yo. datalähteestä käyttäen kaavaa $x \mapsto m(\vec{x}) \pm 0.2$:

- Tietää, että 95% lasketuista estimaateista peittää tuntemattoman parametrin μ (mutta ei tiedä, mitkä niistä)
- Tietää, että 5% lasketuista estimaateista ei peitä μ:tä (mutta ei tiedä, mitkä niistä)

Väliestimaatteja normaalimallista ($\mu = 10, \sigma = 0.5$)



Normaalimallin 99% luottamusväli (tunnettu σ)

Datalähteen normaalimalli:
$X_1, X_2, \ldots$ riippumattomia ja normaalijakautuneita odotusarvona μ (tuntematon) ja keskihajontana σ (tunnettu)

Luottamusvälin määrittäminen:

1. Lasketaan havaitusta datasta keskiarvo $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Määritetään luku $z > 0$, jolle $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.99$
   $\implies z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$
3. Asetetaan parametrin μ luottamusväliksi $m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}}$

Tarkastetaan, että väliestimaatin luottamustaso on 99%.
Datalähteen tuottamalle satunnaisvektorille $\vec{X} = (X_1, \ldots, X_n)$

$$\mathbb{P}\left(|m(\vec{X}) - \mu| \leq z\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}\right| \leq z\right) = \mathbb{P}(|Z| \leq z) = 99\%$$

Normaalimallin odotusarvon estimointi: Yhteenveto

Datalähteen normaalimalli:
$X_1, X_2, \ldots$ riippumattomia ja normaalijakautuneita odotusarvona μ (tuntematon) ja keskihajontana σ (tunnettu)

Parametrin μ suurimman uskottavuuden piste-estimaatti on $m(\vec{x})$
Parametrin μ väliestimaatti on $m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}}$

- 95% luottamustaso, kun $z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
- 99% luottamustaso, kun $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$

Esim. Kun $n = 25$, $\sigma = 0.5$, saadaan väliestimaateiksi:
$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = m(\vec{x}) \pm 0.196$ (95% luottamustasolla)
$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = m(\vec{x}) \pm 0.258$ (99% luottamustasolla)

Käytännön ongelmia:

- Mitä jos σ ei ole ennalta tunnettu?
- Mitä jos datalähde ei noudata normaalimallia?

Normaalimallin odotusarvon estimointi: σ tuntematon

Datalähteen normaalimalli:
$X_1, X_2, \ldots$ riippumattomia ja normaalijakautuneita odotusarvona μ (tuntematon) ja keskihajontana σ (tuntematon)

Asetetaan luottamusväliksi $m(\vec{x}) \pm z \frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$, missä $\mathrm{sd}(\vec{x})$ on havaitun datajoukon keskihajonta.
Datalähteen tuottamalle satunnaisvektorille $\vec{X} = (X_1, \ldots, X_n)$

$$\mathbb{P}\left(|m(\vec{X}) - \mu| \leq z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}\right) = \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\mathrm{sd}(\vec{X})/\sqrt{n}}\right| \leq z\right) = ?$$

Ongelma: $\frac{m(\vec{X})-\mu}{\mathrm{sd}(\vec{X})/\sqrt{n}}$ ei noudata normitettua normaalijakaumaa
Ratkaisu:

- Jos dataa on paljon (n iso), likimain normitettu normaalijakauma
- Jos dataa on vähän, korvataan $\mathrm{sd}(\vec{x})$ otoskeskihajonnalla $\mathrm{sd}_s(\vec{x})$ ja lasketaan $z = -F^{-1}_{t,n-1}(\frac{1-0.99}{2})$ t-jakaumasta

Sisältö

Yleisen stokastisen mallin odotusarvon estimointi

$X_1, X_2, \ldots$ riippumattomia satunnaismuuttujia jostakin jakaumasta, jolla odotusarvo μ (tuntematon)

Parametrin μ piste-estimaatti on $m(\vec{x})$ (ei välttämättä suurimman uskottavuuden estimaatti, mutta harhaton)

Vaikka yksittäiset havainnot eivät tulisi normaalijakaumasta, keskeisen raja-arvolauseen mukaan $m(\vec{X})$ on likimain normaali (jos n riittävän iso).

Likiarvoisen luottamusvälin määrittäminen:

1. Lasketaan havaitusta datasta keskiarvo $m(\vec{x})$ ja keskihajonta $\mathrm{sd}(\vec{x})$
2. Määritetään luku $z > 0$, jolle $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.99$ $\implies$ $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$
3. Asetetaan parametrin μ luottamusväliksi $m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$

Suurille datajoukoille (n iso) pätee $\mathrm{sd}(\vec{X}) \approx \sigma$ ja

$$\mathbb{P}\left(|m(\vec{X})-\mu| \leq z\frac{\mathrm{sd}(\vec{X})}{\sqrt{n}}\right) \approx \mathbb{P}\left(\left|\frac{m(\vec{X})-\mu}{\sigma/\sqrt{n}}\right| \leq z\right) \approx \mathbb{P}(|Z| \leq z) = 99\%.$$

Sisältö

Datalähteen binaarimalli

Datalähteen binaarimalli:
$X_1, X_2, \ldots$ riippumattomia ja $\{0, 1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Parametri p määrittää X_i:n jakauman:

$$\mathbb{E}(X_i) \;=\; 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) \;=\; \mathbb{P}(X_i = 1),$$

joten X_i:n jakauma on

$$f_p(k) \;=\; \begin{cases} 1 - p, & k = 0, \\ p, & k = 1, \\ 0, & \text{muuten.} \end{cases}$$

Tämä on Bernoulli-jakauma parametrina p, merk. $\mathrm{Ber}(p)$ tai myöskin $\mathrm{Bin}(1, p)$.

Esimerkki: Mielipidemittaus

USA:n äänioikeutetuista valittiin satunnaisotannalla $n = 2000$ henkilöä ja heiltä kysyttiin, aikovatko äänestää Trumpia presidentiksi (0=Ei, 1=Kyllä).

Mittaustulos $\vec{X} = (X_1, \ldots, X_{2000})$ noudattaa likimain binaarimallia odotusarvoparametrina p, missä

$$p \;=\; \mathbb{E}(X_i) \;=\; \mathbb{P}(X_i = 1)$$

on Trumpin (tuntematon) kannatus koko populaatiossa.

Tehtävä: Määritä piste-estimaatti ja 95% luottamusväli kannatusosuudelle p.

Edellinen luento: Suurimman uskottavuuden piste-estimaatti $\hat{p} = \hat{p}(\vec{x})$ on ykkösten suhteellinen osuus havaitussa datajoukossa.

Binaarimallin väliestimaatin määrittäminen

Datalähteen binaarimalli:
$X_1, X_2, \ldots$ riippumattomia ja $\{0,1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Koska $p = \mathbb{E}(X_i)$, on tämä erikoistapaus odotusarvoparametrin väliestimoinnista:

1. Lasketaan havaitusta datasta keskiarvo $m(\vec{x})$ ja keskihajonta $\mathrm{sd}(\vec{x})$
2. Määritetään luku $z > 0$, jolle
   $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Asetetaan parametrin p luottamusväliksi $m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$

Käytännön ongelma:

- Yleensä tarkan datajoukon $\vec{x} = (x_1, \ldots, x_n)$ sijaan tiedetään vain datajoukon koko n ja ykkösten suhteellinen osuus $\hat{p}$

Binaarimallin väliestimaatin määrittäminen

Parametrin p luottamusväli on

$$m(\vec{x}) \pm z \frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$$

Miten määritetään luottamusväli, jos tunnetaan vain n ja $\hat{p} = \hat{p}(\vec{x})$?

Binaariarvoiselle datajoukolle:

$$\text{Keskiarvo} \quad m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\#\{i : x_i = 1\}}{n} = \hat{p}$$

$$\text{Varianssi} \quad \mathrm{var}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{p})^2 = \ldots = \hat{p} - \hat{p}^2$$

$$\text{Keskihajonta} \quad \mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})} = \sqrt{\hat{p} - \hat{p}^2}$$

$$\text{Luottamusväli on} \quad \hat{p} \pm z \frac{\sqrt{\hat{p} - \hat{p}^2}}{\sqrt{n}}$$

missä $\hat{p}$ on ykkösten osuus havaitussa datassa.

Binaarimallin väliestimaatti — Yhteenveto

Datalähteen binaarimalli:
$X_1, X_2, \ldots$ riippumattomia ja $\{0,1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Likiarvoisen luottamusvälin (n suuri) määrittäminen:

1. Lasketaan havaitusta datasta ykkösten suhteellinen osuus $\hat{p} = \hat{p}(\vec{x})$
2. Määritetään luku $z > 0$, jolle
   $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Asetetaan parametrin p luottamusväliksi $\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

Binaarimallin konservatiivinen väliestimaatti

Datalähteen binaarimalli:
$X_1, X_2, \ldots$ riippumattomia ja $\{0,1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Joskus halutaan päätellä luottamusvälin leveys ennen tilastokokeen tekemistä (tai halutaan muuten yleispätevä väli, esim. yksi väli kaikille puolueille).
Konservatiivinen väliestimaatti saadaan korvaamalla $\sqrt{\hat{p}(1-\hat{p})}$ luvulla

$$\max_{\hat{p}\in[0,1]} \sqrt{\hat{p}(1-\hat{p})} \;=\; \sqrt{\frac{1}{2}(1-\frac{1}{2})} \;=\; 0.5.$$

Konservatiivisen likiarvoisen luottamusvälin (n suuri) määrittäminen:

1. Lasketaan havaitusta datasta ykkösten suhteellinen osuus $\hat{p}$
2. Määritetään luku $z > 0$, jolle $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Asetetaan p:n luottamusväliksi $\hat{p} \pm z\frac{0.5}{\sqrt{n}}$

Binaarimallin konservatiivinen väliestimaatti

Datalähteen binaarimalli:
$X_1, X_2, \ldots$ riippumattomia ja $\{0, 1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Parametrin p konservatiivinen likiarvoinen luottamusväli on

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

- 95% luottamustaso, kun $z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
- 99% luottamustaso, kun $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$

Mielipidemittauksen virhemarginaali

Mielipidemittauksissa raportoidaan tyypillisesti virhemarginaali (engl. *margin of error, MOE*) muodossa $\pm 2\%$ tai "2% suuntaansa. Tällä tarkoitetaan luottamusvälin pituuden puolikasta, eli jos piste-estimaatti on $\hat{p}$ ja virhemarginaali on h, niin tarkoitetaan että luottamisväli on

$$[\hat{p} - h,\ \hat{p} + h].$$

Luottamustasoa ei aina ilmoiteta, mutta se on usein 95%. Tämä tarkoittaa, että kun samoilla menetelmillä muodostetaan useita luottamusvälejä (esim. eri kuukausina tai eri puolueille), niin pitkän päälle 95% luottamusväleistä sisältää oikean arvon ja 5% ei sisällä.

Mielipidemittauksen luottamusvälin merkityksestä

Oikea arvo p tarkoittaa tässä vastaavaa lukua, joka saataisiin kysymällä sama kysymys koko väestöltä.
Toisin sanoen luottamusväli mittaa vain ns. otantavirhettä eli satunnaisotannasta aiheutuvan virheen vaikutusta estimaattiin $\hat{p}$.

Jos tarkoitus olikin mitata

- väestön "todellista" mielipidettä ($\neq$ kysymykseen annettu vastaus)
- miten väestö todellisuudessa äänestäisi tällä hetkellä
- miten väestö tulee äänestämään 1 kk kuluttua

niin puhutaan eri luvuista. Luottamusväli ei vastaa näihin kysymyksiin.

Ensi viikolla puhutaan bayesläisestä tilastollisesta päättelystä...