MS-A0504 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

4B Tilastolliset luottamusvälit

Lasse Leskelä

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2018–2019
Periodi IV

Esim. Kahviautomaatti

Haluttiin selvittää, kuinka paljon kahviautomaatti keskimäärin laskee kuppiin kahvia. 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 $m(\vec{x}) = 10.03$.

Onko kahviautomaatin valuttamien kahvimäärien todellinen keskiarvo μ lähellä lukua 10.03?

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

Stokastinen malli on tarkka, kun:

- Havaitut alkiot on valittu tasaisen satunnaisesti ja riippumattomasti.
- Havaittujen alkioiden lukumäärä on pieni suhteessa populaation kokoon.

Pienet ja isot kirjaimet

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. (X_1, X_2, X_3) ovat toisistaan riippumattomia normaalijakautuneita satunnaismuuttujia odotusarvona μ ja keskihajontana σ

Datalähteen stokastisen mallin soveltaminen

Ongelma

Otantatutkimuksessa on havaittu arvot $(x_1, \ldots, x_n)$.
Miten voidaan havainnoista päätellä tutkittavan suureen (tuntematon) jakauma koko populaatiossa?

Ratkaisu

Tehdään arvaus, että jakauma on $f(x)$.

Jos arvaus on (likimain) oikea, niin otannan tulosta voidaan mallintaa satunnaismuuttujilla $(X_1, \ldots, X_n)$, jotka ovat toisistaan riippumattomat ja noudattavat (likimain) jakaumaa $f(x)$.

Stokastiikan menetelmillä johdetaan tn, että $(X_1, \ldots, X_n)$ saa (likimain) arvon $(x_1, \ldots, x_n)$.

Jos saatu tn ≈ 0, hylätään arvaus todennäköisin syin.

Datajoukon ja stokastisen mallin tunnusluvut

Stokastiikan menetelmillä johdetaan tn, että $(X_1, \ldots, X_n)$ saa (likimain) arvon $(x_1, \ldots, x_n)$.

- Lasketaan tunnusluku $g(x_1, \ldots, x_n)$ datasta
- Tutkitaan, millä tn:llä satunnaisluku $g(X_1, \ldots, X_n)$ on likimain $g(x_1, \ldots, x_n)$

Tunnusluku on funktio $g : \mathbb{R}^n \to \mathbb{R}$.
(Idea: "sääntö, jolla n havainnon aineistosta lasketaan yksi luku")

Esim

- 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 $\mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})}$

Stokastisen mallin keskiarvo

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:

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

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

Datajoukon ja stokastisen mallin keskiarvot

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

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

$$\mathbb{E}(m(\vec{x})) = m(\vec{x})$$

$$\mathrm{SD}(m(\vec{x})) = 0$$

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

$$m(\vec{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}(m(\vec{X})) = \int_{-\infty}^{\infty} x\, f(x) dx = \mu$$

$$\mathrm{SD}(m(\vec{X})) = \frac{1}{\sqrt{n}} \sigma = \frac{1}{\sqrt{n}} \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx}$$

Yllä μ ja σ ovat hypoteettisen jakauman $f(x)$ odotusarvo ja keskihajonta (jotka lasketaan matemaattisesti, datasta riippumatta).

Stokastisen mallin keskiarvo $m(\vec{X})$ on satunnaisluku, jonka odotusarvo on μ ja keskihajonta $\sigma/\sqrt{n}$.

Yleisen stokastisen mallin keskiarvo

Stokastinen malli: $X_1, \ldots, X_n$ riippumattomia satunnaislukuja odotusarvona μ ja keskihajontana σ

Satunnaismuuttujan $m(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ odotusarvo on

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \mu$$

ja keskihajonta

$$\mathrm{SD}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{SD}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\mathrm{SD}(X_i)^2} = \sigma/\sqrt{n}.$$

Normitetun virheen odotusarvo ja keskihajonta ovat

$$\mathbb{E}\left[\frac{m(X)-\mu}{\sigma/\sqrt{n}}\right] = 0, \qquad \mathrm{SD}\left[\frac{m(X)-\mu}{\sigma/\sqrt{n}}\right] = 1.$$

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.

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

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

Kuinka tarkka tämä estimaatti on?
Millä tn 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 normitettua normaalijakaumaa.

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:

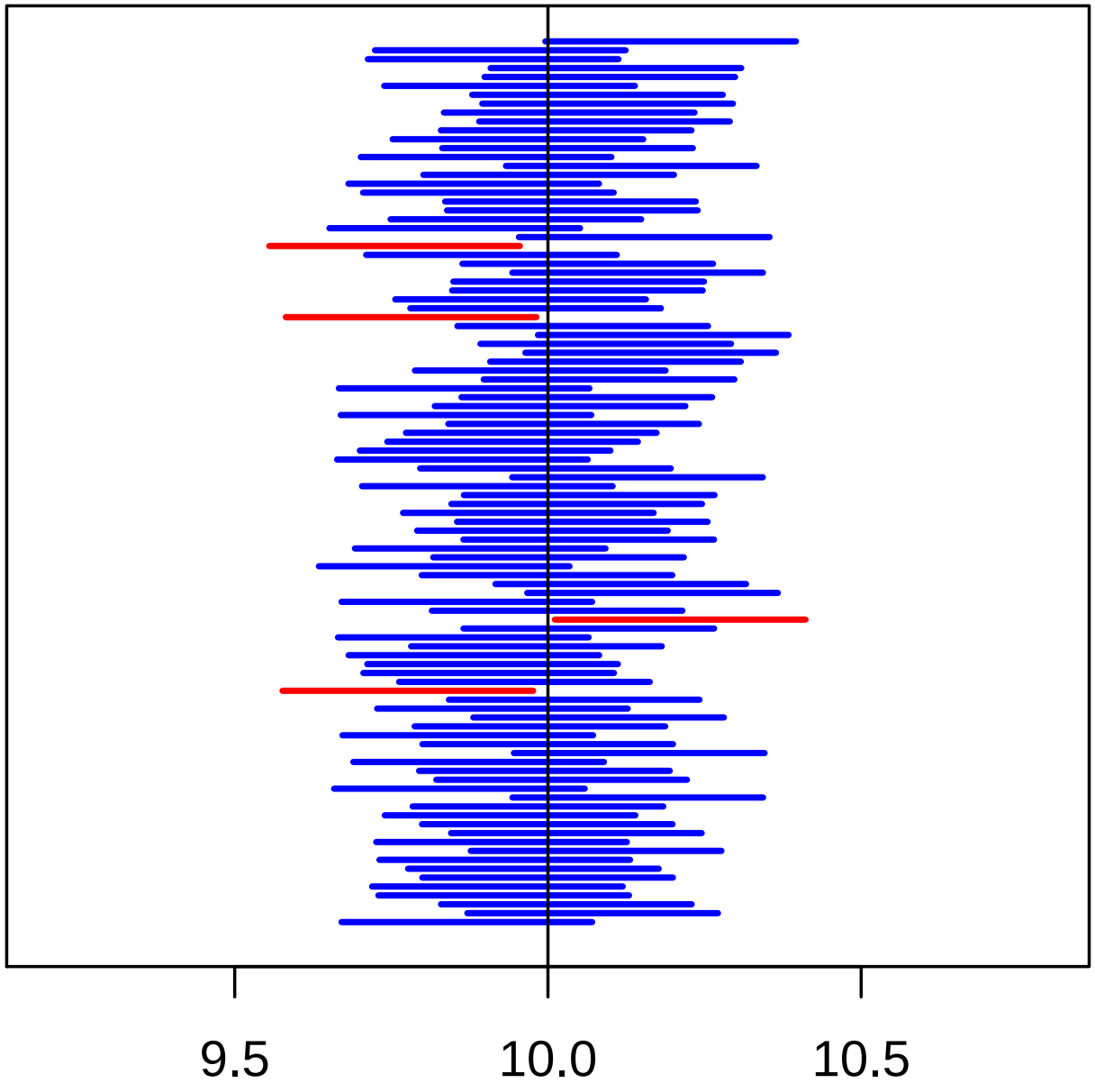
$$\mathbb{P}(\mu \in [m(\vec{X}) - 0.2, m(\vec{X}) + 0.2]) \;=\; 95\%.$$

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

Yleinen stokastinen malli:
$X_1, X_2, \ldots$ riippumattomia odotusarvona μ (tuntematon), jakauma yleinen

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

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:

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

joten X_i:n jakauma on

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

Tämä on Bernoulli-jakauma parametrina p.

Esimerkki: Mielipidemittaus

Usan 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:

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

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

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

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

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.
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 $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\Longrightarrow \; z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
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

> *Helsingin Sanomat kertoi 2. 1. 2008 teettämästään haastattelututkimuksesta, jonka mukaan 78 prosenttia pääkaupunkiseudun 18 vuotta täyttäneistä asukkaista on sitä mieltä, että roskaamisesta tulisi voida sakottaa. Vastaajia oli 1 000, ja virhemarginaaliksi ilmoitettiin 3 prosenttiyksikköä suuntaansa. Se merkitsee, että ~~95 prosentin todennäköisyydellä tätä mieltä on 75–81 prosenttia pääkaupunkiseudun kansalaisista.~~* [Tiede 3/2008]

Väärin. Parametrin p (roskaamisen sakottamisen kannattajat) konservatiivinen likiarvoinen väliestimaatti 95% luottamustasolla on

$$\hat{p}(\vec{x}) \pm 1.96 \frac{0.5}{\sqrt{1000}} \;=\; 0.78 \pm 0.03.$$

Oikea tulkinta: Jos kysely toistettaisiin uudelle satunnaisotokselle $\vec{X}$, niin 95% tn:llä väli $\hat{p}(\vec{X}) \pm 0.03$ peittäisi tuntemattoman parametrin p.

http://www.tiede.fi/artikkeli/kysy/mika_on_virhemarginaali

Mielipidemittauksen virhemarginaali

> *Virhemarginaali tarkoittaa satunnaisvaihtelusta syntyvää epävarmuustekijää tuloksissa, kun käytetään otostutkimuksia. Mielipidetutkimuksissa ilmoitetaan esimerkiksi, että 95 %:n virhemarginaali 1000 hengen kyselyssä on ±3 %. Tämä tarkoittaa sitä, että jos poimisimme suuren joukon samanlaisia otoksia, 95 otoksessa 100:sta otoksesta laskettu arvo on näiden marginaalien rajoissa ja vain 5 % otoksista niiden ulkopuolella.*
> [Tilastokeskus]

Selvä? Parametrin p konservatiivinen likiarvoinen väliestimaatti 95% luottamustasolla on

$$\hat{p}(\vec{x}) \pm 1.96 \frac{0.5}{\sqrt{1000}} \approx \hat{p}(\vec{x}) \pm 0.03.$$

Selkeämpi tulkinta: Jos poimisimme suuren joukon samanlaisia otoksia, otoksesta laskettu luottamusväli sisältäisi tuntemattoman parametrin arvon 95 otoksessa 100:sta.

http://www.stat.fi/meta/kas/virhemarginaali.html

Mielipidemittauksen virhemarginaali: Yhteenveto

Mielipidemittauksen virhemarginaali tarkoittaa datalähteen binaarimallin tuntemattoman parametrin p (kannattajien osuus) väliestimaattia valitulla luottamustasolla.

- Yleensä unohdetaan mainita, mitä luottamustasoa on käytetty (yleensä 95%).
- Koskaan ei kerrota, mitä kaavaa on käytetty.
- Suurin osa kansantajuisista virhemarginaalin määrittelyistä on virheellisiä tai epäselviä.

Konservatiivisen likiarvoisen 95% luottamustason $\hat{p}(\vec{x}) \pm 1.96 \times \frac{0.5}{\sqrt{n}}$ määrittämiseen riittää tietää n:

- $n = 1000 \implies \hat{p}(\vec{x}) \pm 3\%$
- $n = 2000 \implies \hat{p}(\vec{x}) \pm 2\%$
- $n = 9000 \implies \hat{p}(\vec{x}) \pm 1\%$

Ensi viikolla puhutaan bayesläisestä tilastollisesta päättelystä...