MS-A0504 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

5B Bayesläiset piste- ja väliestimaatit

Lasse Leskelä

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2018–2019
Periodi IV

Sisältö

Posterijakauman tulkinta: Piste-estimaatit

Binaarimallin estimointi

Binaarimallin estimointi: Frekventistinen tapa

Binaarimallin estimointi: Bayesläinen tapa

Yhteenveto estimoinnista

Posteriorijakauman estimaatit

Miten päätellään tuntemattoman parametrin θ estimaatti posteriorijakaumasta $p_1(\theta \mid \vec{x})$?

1. Posteriorijakauman moodi $\hat{\theta}$:

$$p_1(\hat{\theta} \mid \vec{x}) \;=\; \max_{\theta} p_1(\theta \mid \vec{x})$$

2. Posteriorijakauman odotusarvo $\hat{\theta}$:

$$\hat{\theta} \;=\; \sum_{\theta} \theta p_1(\theta \mid \vec{x})$$

3. Raportoidaan koko posteriorijakauma

Vaihtoehto 1. usein helpompi määrittää: (ei tarvitse tuntea posteriorijakauman normitusvakiota)

Esimerkki: Tasajakauman ylärajan estimointi

Tuntemattoman välin $\{1, 2, \ldots, \theta\}$ tasajakaumaa noudattavasta datalähteestä on havaittu $x_1 = 21$, $x_2 = 7$ ja $x_3 = 22$. Mikä on paras arvaus (estimaatti) tuntemattoman parametrin θ arvolle?

Datalähteen uskottavuusfunktio:

$$f(x_i \,|\, \theta) \;=\; \begin{cases} \frac{1}{\theta}, & 1 \leq x_i \leq \theta, \\ 0, & \text{muuten} \end{cases}$$

$$f(\vec{x} \,|\, \theta) \;=\; \begin{cases} \frac{1}{\theta^3}, & 1 \leq x_1, x_2, x_3 \leq \theta, \\ 0, & \text{muuten} \end{cases}$$

Suurimman uskottavuuden estimaatti: $\hat{\theta}(\vec{x}) = \max(x_1, x_2, x_3) = 22$.

Entä jos meillä on ennakkoon muodostettu näkemys, että tuntemattoman parametrin arvo on todennäköisesti lähellä arvoa 30?

Tasajakauman yläraja: bayeslainen estimointi

Tuntemattoman välin $\{1, 2, \ldots, \theta\}$ tasajakaumaa noudattavasta datalähteestä on havaittu $x_1 = 21$, $x_2 = 7$ ja $x_3 = 22$.

Ennakkotietämys: Parametri on todennäköisesti lähellä arvoa 30.

Tulkitaan tuntematon parametri satunnaismuuttujana, jonka odotusarvo on 30?

- Miten valitaan priorijakauma?
- Valitaan yksinkertaisin diskreetti jakauma, jonka odotusarvo on 30 ja jolla on positiivinen tn saada mikä tahansa pos. kokonaisluku.
- Poisson-jakauma parametrina $\lambda = 30$:

$$p_0(\theta) \;=\; e^{-\lambda}\frac{\lambda^\theta}{\theta!}, \quad \theta = 0, 1, 2, \ldots$$

- Uskottavuusfunktio

$$f(\vec{x}\,|\,\theta) \;=\; \begin{cases} \frac{1}{\theta^3}, & 1 \le x_1, x_2, x_3 \le \theta, \\ 0, & \text{muuten} \end{cases}$$

Tasajakauman ylärajan estimointi: bayeslainen estimointi

Havaittu data: $\vec{x} = (21, 7, 22)$

$$\text{Priori} \qquad p_0(\theta) \;=\; e^{-30}\frac{30^\theta}{\theta!}, \qquad \theta = 0, 1, 2, \ldots$$

$$\text{Uskottavuus} \qquad f(\vec{x}\,|\,\theta) \;=\; \begin{cases} \frac{1}{\theta^3}, & \theta \geq 22, \\ 0, & \text{muuten} \end{cases}$$

Posteriori lasketaan päivityskaavasta

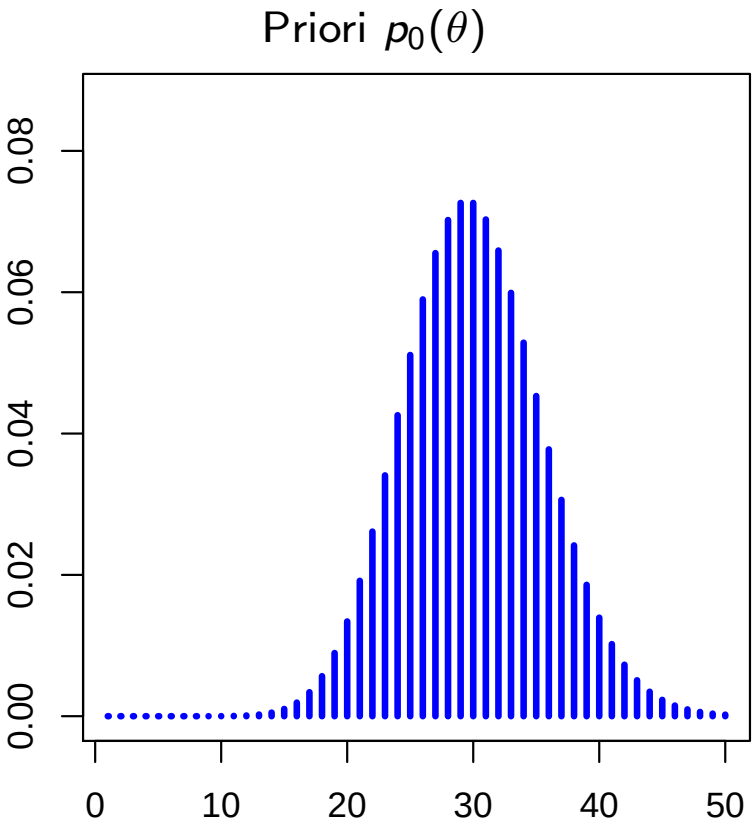
$$p_1(\theta\,|\,\vec{x}) \;=\; \frac{p_0(\theta) f(\vec{x}\,|\,\theta)}{\sum_{\theta'} p_0(\theta') f(\vec{x}\,|\,\theta')} \;=\; \begin{cases} c^{-1} e^{-30}\frac{30^\theta}{\theta!}\frac{1}{\theta^3}, & \theta \geq 22, \\ 0, & \text{muuten}, \end{cases}$$

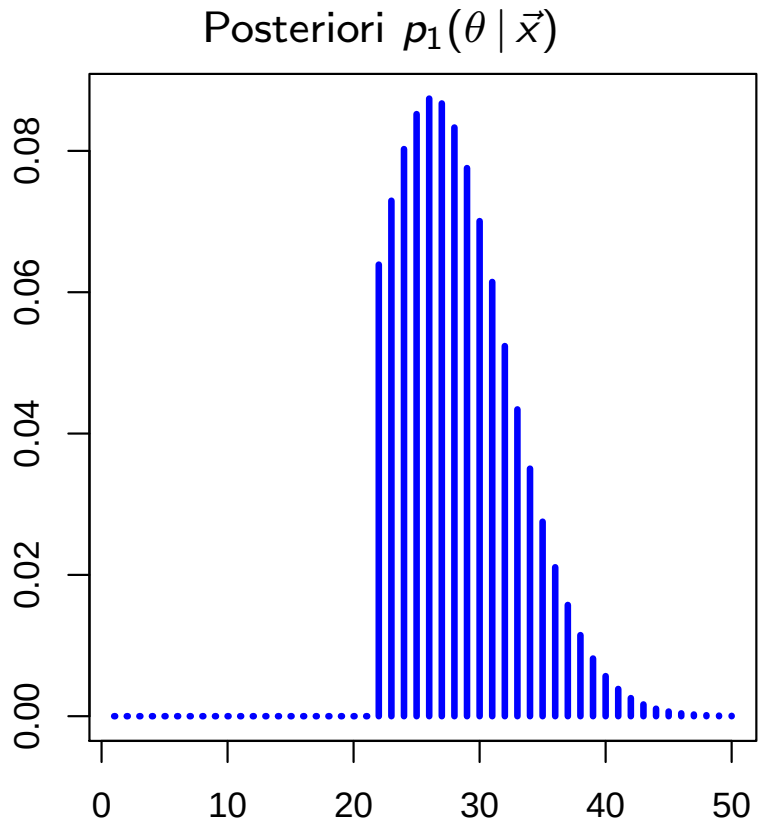
missä normitusvakio $c = \sum_{\theta'} p_0(\theta') f(\vec{x}\,|\,\theta')$ ei riipu θ:sta.
Suurimman posterioritodennäköisyyden estimaatti on $\hat{\theta}$, joka maksimoi funktion $\theta \mapsto \frac{30^\theta}{\theta!}\frac{1}{\theta^3}$ joukossa $\theta \geq 22$.
Maksimin voi etsiä kokeilemalla tai piirtämällä ko. funktio.

Tasajakauman yläraja: Priori ja posteriori

Data: $\vec{x} = (21, 7, 22)$

Priori: Poisson-jakauma odotusarvona $\lambda = 30$





Suurimman posterioritodennäköisyyden estimaatti: $\hat{\theta}(\vec{x}) = 26$

Sisältö

Posterijakauman tulkinta: Piste-estimaatit

Binaarimallin estimointi

Binaarimallin estimointi: Frekventistinen tapa

Binaarimallin estimointi: Bayesläinen tapa

Yhteenveto estimoinnista

Datalähteen binaarimalli

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

Parametri p määrittää datalähteen jakauman,
X_i:n pistetodennäköisyyfunktio on

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

Tämä on Bernoulli-jakauma parametrina p.

Esimerkki: Mielipidemittaus

Usan äänioikeutetuista valittiin satunnaisotannalla $n = 200$ henkilöä ja heiltä kysyttiin, aikovatko äänestää Trumpia presidentiksi (0=Ei, 1=Kyllä). 70 vastasi kyllä.

Mittaustulos $X = (X_1, \ldots, X_{200})$ 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 luottamusväli kannatusosuudelle p.

Sisältö

Posterijakauman tulkinta: Piste-estimaatit

Binaarimallin estimointi

Binaarimallin estimointi: Frekventistinen tapa

Binaarimallin estimointi: Bayesläinen tapa

Yhteenveto estimoinnista

Estimaatti: Frekventistinen tapa

Parametrin p suurimman uskottavuuden estimaatti on

$$\hat{p} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; \frac{\#\{i : x_i = 1\}}{n}$$

eli ykkösten suhteellinen osuus datajoukossa $\vec{x}$.

Likiarvoinen (n suuri) 95% luottamusväli on

$$\hat{p} \pm z\,\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}},$$

- $z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$ on luku, jolle $\mathbb{P}(|Z| \leq z) = 0.95$

Entä jos käytössä on ennakkotietoa kannatusosuudesta p?

Esim. aiempien mielipidemittausten mukaan arvioidaan, että kannatus todennäköisesti on lähellä lukua 0.4.

Sisältö

Posterijakauman tulkinta: Piste-estimaatit

Binaarimallin estimointi

Binaarimallin estimointi: Frekventistinen tapa

Binaarimallin estimointi: Bayesläinen tapa

Yhteenveto estimoinnista

Bayesläinen estimointi

Tulkitaan tuntematon kannatusosuus satunnaismuuttujaksi Θ, jonka priorijakauma mallintaa ennakkotietämystä kannatuksesta.

Miten priorijakauma valitaan?

- Uskotaan, että Θ on todennäköisesti 0.4?
- Uskotaan, että $\Theta \in [0.3, 0.5]$ 95% todennäköisyydellä.

Jos valitaan välin $[0.3 - 1/190, 0.5 + 1/190]$ tasajakauma, saadaan

$$\mathbb{P}(\Theta \in [0.3, 0.5]) \;=\; \frac{0.2}{0.2 + 2/190} \;=\; 95\%.$$

Onko tämä hyvä priorijakauma?
Tuskin, sillä pisteissä, missä priorijakauma on 0, on myös posteriorijakauma väistämättä 0.

Priorijakauman valitseminen

Miten priorijakauma valitaan, kun uskotaan että $\Theta \in [0.3, 0.5]$ todennäköisyydellä 95%?

Keksi jatkuvan välin $[0, 1]$ jakauma $f_0(t)$, jonka

- odotusarvo $\int t\, f(t) dt = 0.4$
- tiheysfunktio $f_0(t) > 0$ kaikilla $t \in (0, 1)$
- $\int_{0.3}^{0.5} f_0(t) dt \approx 0.95$

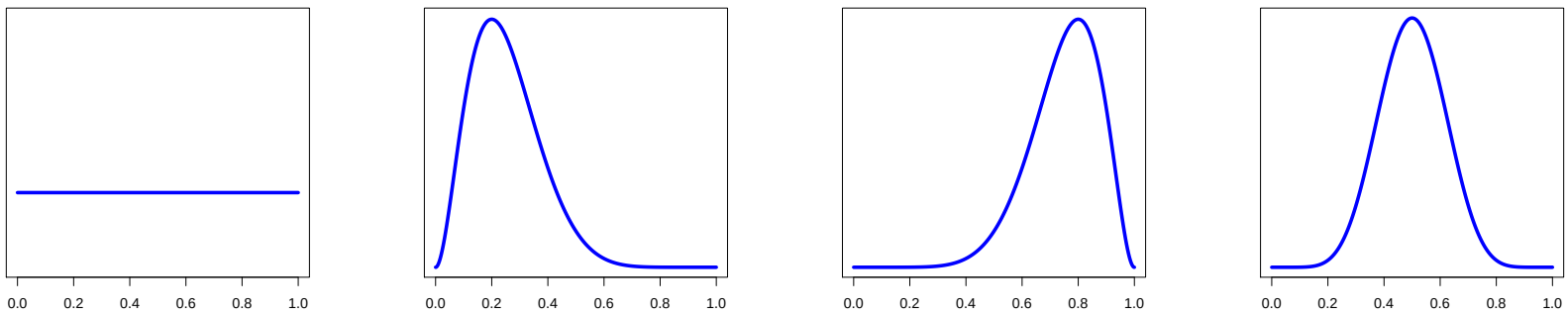
Beta-jakauma

Beta(a, b)-jakauman parametreina $a > 0$ ja $b > 0$ tiheysfunktio on

$$f(\theta) = \begin{cases} c\,\theta^{a-1}(1-\theta)^{b-1}, & \text{kun } \theta \in [0,1], \\ 0, & \text{muuten,} \end{cases}$$

normitusvakiona $c = \frac{(a+b-1)!}{(a-1)!(b-1)!}$.

Beta$(1,1)$ Beta$(3,9)$ Beta$(9,3)$ Beta$(9,9)$



- Arvojoukko $= [0,1]$
- Odotusarvo $\mu = \frac{a}{a+b}$ ja keskihajonta $\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}}$
- Kertymäfunktiota ei tunneta suljetussa muodossa

`dbeta(theta,a,b); pbeta(theta,a,b)`

Priorijakauman valitseminen

Miten priorijakauma valitaan, kun uskotaan että $\Theta \in [0.3, 0.5]$ todennäköisyydellä 95%?

Valitaan odotusarvoksi $\mu = 0.4$ ja keskihajonnaksi ?

- Normitetulle normaalijakaumalle $Z = 0 \pm 2$ tn:llä 95%
- Yleiselle normaalijakaumalle $Z = \mu \pm 2\sigma$ tn:llä 95%
- Jos Beta-jakauman hajonta normaalin kaltaista, niin Beta-jakaumalle $X = \mu \pm 2\sigma$ suurin piirtein tn:llä 95%
- Valitaan $\sigma = 0.1/2 \approx 0.05$.

Ratkaistaan Beta(a, b)-jakauman parametrit:

$$\mu = \frac{a}{a+b} \qquad a = \mu\left(1 + \frac{\mu(1-\mu)}{\sigma^2}\right) = 38$$

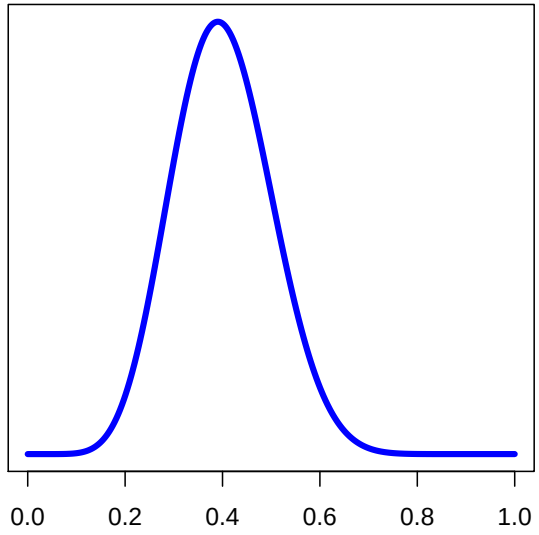
$$\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}} \qquad b = (1-\mu)\left(1 + \frac{\mu(1-\mu)}{\sigma^2}\right) = 57$$

Priorijakauman valitseminen

Miten priorijakauma valitaan, kun uskotaan että $\Theta \in [0.3, 0.5]$ todennäköisyydellä 95%?

Kokeillaan Beta$(38, 57)$-jakaumaa:

- Odotusarvo $\mu = 0.4$
- Keskihajonta $\sigma = 0.05$



Tällöin

$$\mathbb{P}(\Theta \in [0.3, 0.5]) = F_{38,57}(0.5) - F_{38,53}(0.3) = 0.9551861.$$

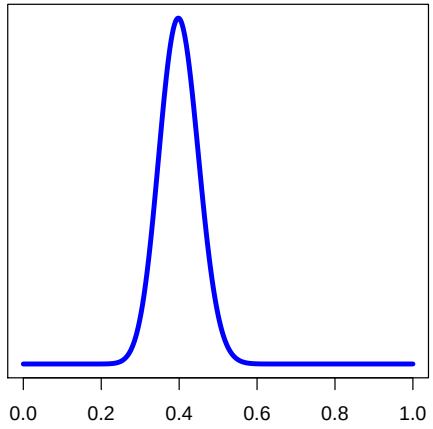
```
pbeta(0.5,38,57)-pbeta(0.3,38,57)
```

Posteriorijakauman määrittäminen

Havaitaan $n = 200$ alkion datajoukko, jossa 70 ykköstä ja 130 nollaa (ykkösten osuus $= 35\%$) .

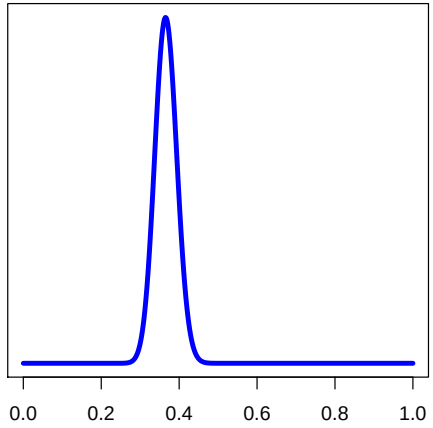
Priorijakauma
Beta$(38, 57)$:

- Odotusarvo $\mu_0 = 0.4$
- Keskihajonta $\sigma_0 = 0.05$



Posteriorijakauma
Beta$(38 + 70, 57 + 130) =$ Beta$(108, 187)$:

- Odotusarvo $\mu_1 = \frac{108}{108+187} = 0.366$
- Keskihajonta $\sigma_1 = 0.028$



Bayeslainen piste-estimaatti $=$ posteriorijakauman odotusarvo 0.366

Bayeslainen väliestimaatti

Etsitään piste-estimaatin $\mu_1 = 0.366$ ympäriltä väli, johon posteriorijakaumaa noudattava Θ kuuluu tn:llä 95%.
Miten?
Jos Beta$(108, 187)$ hajonnaltaan normaalin kaltainen, niin voidaan kokeilla väliä

$$\mu_1 \pm 2\sigma_1 \;=\; 0.366 \pm 2 \times 0.028 \;=\; 0.366 \pm 0.056$$

Tällöin

$$\mathbb{P}(\Theta \in [0.310, 422] \,|\, \vec{x}) \;\approx\; 95.47\%$$

Johtopäätös:
Havaitun datan (70 ykköstä, $n = 200$) valossa tuntemattoman parametrin Θ ehdollinen todennäköisyys kuulua välille 0.366 ± 0.056 on noin 95%.

Sisältö

Posterijakauman tulkinta: Piste-estimaatit

Binaarimallin estimointi

Binaarimallin estimointi: Frekventistinen tapa

Binaarimallin estimointi: Bayesläinen tapa

Yhteenveto estimoinnista

Yhteenveto

Frekventistinen

- Data: 70 ykköstä, 130 nollaa
- Parametrin arvoihin ei liitetä todennäköisyyksiä
- Parametrilla ei ole priorijakaumaa
- Parametrilla ei ole posteriorijakaumaa
- Piste-estimaatti on suurimman uskottavuuden estimaatti $\hat{p} = \frac{70}{200} = 0.350$
- 95% luottamustason luottamusväli $\hat{p} \pm 1.96 \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = 0.350 \pm 0.067$
- 95% samalla menetelmällä estimoiduista luottamusväleistä peittää parametrin

Bayeslainen

- Data: 70 ykköstä, 130 nollaa
- Parametrin arvoihin liitetään subj. todennäköisyydet
- Priorijak. Beta$(38, 57)$ $\mu_0 = 0.400$, $\sigma_0 = 0.050$
- Posteriorijak. Beta$(108, 187)$ $\mu_1 = 0.366$, $\sigma_1 = 0.028$
- Piste-estimaatti on posteriorijakauman odotusarvo $\mu_1 = 0.366$.
- Väliestimaatti on 95% posteriorijakauman väli $\mu_1 \pm 2\sigma_1 = 0.366 \pm 0.056$
- 95% todennäköisyydellä parametri kuuluu välille 0.366 ± 0.056

Ensi viikolla jatketaan tilastollisen merkitsevyyden testaamisesta...