Sovellettu todennäköisyyslaskenta B

Antti Rasila

30. lokakuuta 2007

1. Otos ja otosjakaumat (jatkoa)
   - Frekvenssi ja suhteellinen frekvenssi
   - Frekvenssien odotusarvo ja varianssi
   - Suhteellisen frekvenssin normaaliapproksimaatio
2. Estimointi
   - Estimaatti ja estimaattori
   - Hyvä estimaattori
   - Estimaattorin keskineliövirhe ja tarkkuus
   - Piste-estimointi ja väliestimointi
3. Luottamusvälin määritys
   - Normaalijakautuneen estimaattorin määräämä luottamusväli, kun varianssi tunnetaan
   - Normaalijakautuneen satunnaismuuttujan odotusarvon luottamusväli, kun varianssi tunnetaan
   - Normaalijakauman odotusarvon luottamusväli
   - Normaalijakauman varianssin luottamusväli

Frekvenssi ja suhteellinen frekvenssi

- Olkoon A jokin otosavaruuden S alkioiden ominaisuus (eli S:n osajoukko).
- Poimitaan otosavaruudesta yksinkertainen satunnaisotos, jonka koko on n.
- Ominaisuuden A omaavien alkioden lukumäärää otoksessa merkitään f:llä kutsutaan ominaisuuden A *frekvenssiksi*.
- Ominaisuuden A *suhteellinen frekvenssi* $\hat{p}$ määritellään:

$$\hat{p} = \frac{f}{n}$$

Frekvenssien odotusarvo ja varianssi

- Frekvenssi f noudattaa eksaktisti *binomijakaumaa* parametrein n ja $Pr(A) = p$:

$$f \sim Bin(n, p)$$

- Frekvenssin f odotusarvo ja varianssi ovat siis:

$$E(f) = np$$
$$Var(f) = npq,$$

missä $q = 1 - p$.

- Vastaavasti suhteellisen frekvenssin $\hat{p}$ odotusarvo ja varianssi ovat:

$$E(\hat{p}) = p$$
$$Var(\hat{p}) = \frac{pq}{n}$$

Suhteellisen frekvenssin normaaliapproksimaatio

- Suhteellinen frekvenssi $\hat{p}$ noudattaa suurissa otoksissa approksimatiivisesti normaalijakaumaa:

$$\hat{p} \sim_a N\left(p, \frac{pq}{n}\right)$$

- Vastaavasti, standardoitu satunnaismuuttuja

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}},$$

noudattaa suurissa otoksissa standardoitua normaalijakumaa:

$$Z \sim_a N(0, 1)$$

Estimaatti ja estimaattori (1/2)

- Tilastollisen tutkimuksen tärkeimpiä osatehtäviä on *estimoida* eli arvioida tutkimuksen kohteena olevaa ilmiötä.
- Ilmiötä koskevat havainnot generoineen prosessin mallina käytettävän todennäköisyysjakauman tuntemattomat parametrit päätellään ilmiötä koskevien havaintojen perusteella.
- Havaintojen funktiota, joka tuottaa *estimaatteja* parametrin todelliselle arvolle, kutsutaan parametrin *estimaattoriksi*.

Estimaatti ja estimaattori (1/2)

- Oletetaan, että satunnaismuuttuja X noudattaa todennäköisyysjakaumaa, jonka pistetodennäköisyys- tai tiheysfunktio $f(x;\theta)$ riippuu *parametrista* θ.
- *Parametrin* θ *estimoimiseen* käytetään *havaintojen* $X_1, X_2, \ldots, X_n$ funktiota, eli *tunnuslukua*

$$T = g(X_1, X_2, \ldots, X_n) = \hat{\theta}$$

- Funktiota T kutsutaan *parametrin* θ *estimaattoriksi*.
- Havaintoarvoista $x_1, x_2, \ldots, x_n$ laskettua arvoa

$$t = g(x_1, x_2, \ldots, x_n)$$

kutsutaan *parametrin* θ *estimaatiksi*.

Hyvä estimaattori

- Todennäköisyysjakauman parametreille on tavallisesti tarjolla useita vaihtoehtoisia estimaattoreita.
- Seuraavat kriteerit täyttävä estimaattori tuottaa järkeviä arvoja estimoitavalle parametrille:
  - *Harhattomuus*
  - *Tyhjentävyys*
  - *Tehokkuus*
  - *Tarkentuvuus*

Harhattomuus ja harha

- Olkoon $\hat{\theta}$ parametrin θ estimaattori.
- Estimaattori $\hat{\theta}$ on *harhaton* parametrille θ, jos sen odotusarvo yhtyy parametrin arvoon, eli
$$\mathrm{E}(\hat{\theta}) = \theta.$$
- Estimaattorin $\hat{\theta}$ *harha* on erotus
$$\mathrm{Bias}(\tilde{\theta}) = \theta - \mathrm{E}(\hat{\theta}).$$
- Toisin sanoen, jos esitimaattori $\hat{\theta}$ on harhaton parametrille θ, niin $\mathrm{Bias}(\hat{\theta}) = 0$.

Tyhjentävyys

- Parametrin θ estimaattori $\hat{\theta}$ on tyhjentävä, jos se käyttää kaiken otoksessa olevan informaation.
- Tämä tarkoittaa sitä, että jos $\hat{\theta}_1$ on mikä tahansa parametrin θ estimaattori, niin sen jakauma ei riipu θ:sta, jos $\hat{\theta}$ on tunnettu.

Tehokkuus

- Oletetaan, että $\hat{\theta}_1$ ja $\hat{\theta}_2$ ovat parametrin θ estimaattoreja.
- Estimaattori $\hat{\theta}_1$ on *tehokkaampi* kuin estimaattori $\hat{\theta}_2$, jos

$$\mathrm{D}^2(\hat{\theta}_1) < \mathrm{D}^2(\hat{\theta}_2).$$

- Parametrin θ estimaattori $\hat{\theta}$ on *täystehokas*, jos sen varianssi on pienempi kuin minkä tahansa muun saman parametrin estimaattorin.

Tarkentuvuus

- Estimaattori $\hat{\theta}$ on *tarkentuva* parametrille θ, jos se *konvergoi melkein varmasti* kohti parametrin oikeaa arvoa, kun otoskoko kasvaa rajatta.
- Tämä tarkoittaa sitä, että

$$Pr(T_n \to 0) = 1, \text{ kun } n \to \infty.$$

Estimaattorin keskineliövirhe ja tarkkuus

- Olkoon $\hat{\theta}$ parametrin θ estimaattori. Tällöin $\hat{\theta}$:n *keskineliövirhe* on

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{D}^2(\hat{\theta}) + \big[\mathrm{Bias}(\hat{\theta})\big]^2.$$

- Jos $\hat{\theta}$ on parametrin θ harhaton estimaattori, eli

$$\mathrm{Bias}(\hat{\theta}) = \theta - \mathrm{E}(\hat{\theta}) = 0,$$

niin

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{D}^2(\hat{\theta}) = 0.$$

- Estimaattoria sanotaan *tarkaksi*, jos se on harhaton ja sen varianssi on pieni.

Piste-estimointi ja väliestimointi

- Todennäköisyysjakauman *parametrin arvon estimointia* kutsutaan *piste-estimoinniksi*.
- Parametrin estimaattiin on aina syytä liittää *luottamusväliksi* kutsuttu väli. Parametrin todellinen arvo sijoittuu luottamusvälille soveltajan valittavissa olevalla todennäköisyydellä.
- Täsmällisemmpin väli $[a, b]$ on parametrin θ luottamusväli luottamustasolla $1 - \alpha$, jos

$$Pr(a \leq \theta \leq b) \geq 1 - \alpha.$$

- Tavallisimmat luottamusvälit ovat 95%, 99% ja 99.9%. Näitä vastaavat α:n arvot ovat $0.05, 0.01$ ja 0.001.
- Luottamusvälin määräämistä kutsutaan *väliestimoinniksi*.

Luottamusvälin määritys 1/2

- Tehdään seuraavat oletukset:
  - Satunnaismuuttuja X noudattaa jakaumaa $f(x;\theta)$.
  - $X_1, X_2, \ldots, X_n$ on yksinkertainen satunnaisotos jakaumasta $f(x;\theta)$.
  - $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ on θ:n estimaattori.
- Valitaan *luottamustaso* $1 - \alpha$ ja määrätään satunnaismuuttujat

$$A = A(X_1, X_2, \ldots, X_n)$$
$$Y = Y(X_1, X_2, \ldots, X_n)$$

siten että

$$Pr(\hat{\theta} - A \leq \theta) = \frac{\alpha}{2}$$
$$Pr(\hat{\theta} + Y \geq \theta) = \frac{\alpha}{2}$$

Luottamusvälin määritys 2/2

- Nyt todennäköisyys

$$Pr(\hat{\theta} - A \leq \theta \leq \hat{\theta} + Y) = 1 - \alpha$$

ja väli

$$(\hat{\theta} - A, \hat{\theta} + Y)$$

on θ:n *luottamusväli luottamustasolla* $(1 - \alpha)$.

- Jos $\hat{\theta}$:n jakauma on *symmetrinen*, pätee $A = Y$ luottamusväli on muotoa

$$(\hat{\theta} - A, \hat{\theta} + A).$$

Normaalijakautuneen estimaattorin määräämä luottamusväli, kun varianssi tunnetaan

- Oletetaan, että satunnaismuuttuja $\hat{\theta} \sim N(\mu, \sigma^2)$ on parametrin θ estimaattori.
- Tällöin standardoitu satunnaismuuttuja

$$Z = \frac{\hat{\theta} - \mu}{\sigma} \sim N(0, 1).$$

- Tällöin

$$Pr(-z_{\alpha/2} \leq \frac{\hat{\theta} - \mu}{\sigma} \leq z_{\alpha/2}) = 1 - \alpha.$$

- Parametrille θ saadaan siis $(1 - \alpha)$-luottamusväliksi

$$\hat{\theta} - z_{\alpha/2}\sigma \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma. \quad (1)$$

Normaalijakautuneen satunnaismuuttujan odotusarvon luottamusväli, kun varianssi tunnetaan

- Olkoon $X_1, \ldots, X_n$ yksinkertainen satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$. Oletetaan, että σ tunnetaan, mutta μ on tuntematon.
- Tällöin havaintojen aritmeettinen keskiarvo $\overline{X}$ noudattaa eksaktisti normaalijakaumaa:

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

- Kaavasta (1) odotusarvon μ $(1-\alpha)$-luottamusväliksi saadaan

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \qquad (2)$$

Normaalijakauman odotusarvon luottamusväli

- Olkoon havainnot $X_1, \ldots, X_n$ yksinkertainen satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$ ja olkoon

  $\overline{X}$ = havaintojen aritmeettinen keskiarvo
  s^2 = havaintojen harhaton otosvarianssi
  n = havaintojen lukumäärä
  $t_{\alpha/2}$ = t-jakauman arvo merkitsevyystasolla $\alpha/2$ ja vapausasteilla $(n-1)$.

- Normaalijakauman *odotusarvoparametrin* μ *luottamusväli luottamustasolla* $(1-\alpha)$ on muotoa

$$\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{3}$$

Normaalijakauman varianssin luottamusväli

- Olkoon havainnot $X_1, \ldots, X_n$ yksinkertainen satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$ ja olkoon

$$
\begin{aligned}
s^2 &= \text{havaintojen harhaton otosvarianssi} \\
n &= \text{havaintojen lukumäärä} \\
\chi^2_{1-\alpha/2} \text{ ja } \chi^2_{\alpha/2} &= \chi^2\text{-jakauman arvot merkitsevyystasoilla} \\
&\quad 1-\alpha/2 \text{ ja } \alpha/2 \\
&\quad \text{ja vapausasteilla } (n-1).
\end{aligned}
$$

- Normaalijakauman *varianssiparametrin* σ^2 *luottamusväli luottamustasolla* $(1-\alpha)$ on muotoa

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

Esimerkki (Niemi) 1/3

- Paperilaadun neliöpainon määrittämiseksi punnitaan 16 arkkia paperia ja muunnetaan saadut painot neliöpainoiksi (g/m^2):

78.90	80.39	78.55	81.81	78.05	79.84	81.85	80.83
79.43	81.31	78.44	79.52	80.14	81.37	82.80	79.54

- Oletetaan neliöpaino normaalijakautuneeksi. Muodostetaan odotusarvolle μ luottamusvälit luottamustasoilla 0.95 ja 0.99, kun
  (a) tiedetään, että keskihajonta $\sigma = 1.3\, g/m^2$, ja
  (b) keskihajonta on tuntematon.
- Otoskeskiarvo on $\bar{x} = 80.11\, g/m^2$.

Esimerkki (Niemi) 2/3

- Tapaus (a): Normaalijakauman kriittiset arvot ovat $z_{0.025} = 1.96$ ja $z_{0.005} = 2.58$.
- Luottamustasolla 0.95 luottamusväliksi saadaan kaavasta (2)

$$\left[80.11 - 1.96 \cdot \frac{1.3}{\sqrt{16}}, 80.11 + 1.96 \cdot \frac{1.3}{\sqrt{16}}\right] = [79.47, 80.75].$$

- Vastaavasti, luottamustasolla 0.99 luottamusväliksi saadaan

$$\left[80.11 - 2.58 \cdot \frac{1.3}{\sqrt{16}}, 80.11 + 2.58 \cdot \frac{1.3}{\sqrt{16}}\right] = [79.27, 80.95].$$

Esimerkki (Niemi) 3/3

- Tapaus (b): Koska keskihajonta on tuntematon, estimoidaan se myös otoksesta: $s = 1.33\, g/m^2$. Otoksen koko on 16, joten luottamusvälin laskemisessa käytetään t-jakaumaa vapausasteilla 15.
- Taulukosta tai tietokoneella t-jakauman kriittisiksi arvoiksi saadaan $t_{0.025} = 2.131$ ja $t_{0.005} = 2.947$.
- Luottamustasolla 0.95 luottamusväliksi saadaan kaavasta (3)

$$\left[80.11 - 2.131 \cdot \frac{1.33}{\sqrt{16}}, 80.11 + 2.131 \cdot \frac{1.33}{\sqrt{16}}\right] = [79.40, 80.82],$$

  ja luottamustasolla 0.99

$$\left[80.11 - 2.947 \cdot \frac{1.33}{\sqrt{16}}, 80.11 + 2.947 \cdot \frac{1.33}{\sqrt{16}}\right] = [79.13, 81.01].$$

- Huomaa, että t-jakauman kriittiset arvot ovat kauempana origosta kuin standardinormaalijakauman vastaavat arvot. Tämä on luonnollista, koska epävarmuustekijöitä on enemmän.