Sovellettu todennäköisyyslaskenta B

Antti Rasila

18. lokakuuta 2007

1. Tilastollinen aineisto

2. Tilastollinen malli
   - Yksinkertainen satunnaisotos

3. Otostunnusluvut ja otosjakaumat
   - Aritmeettisen keskiarvon odotusarvo ja varianssi
   - Otosvarianssin odotusarvo ja varianssi
   - Otosvarianssin otosjakauma, kun otos on normaalijakautunut
   - Jakauman vinous ja huipukkuus
   - Geometrinen- ja harmoninen keskiarvo

Tilastollinen aineisto

- *Tilastollinen aineisto* koostuu tutkimuksen kohteita kuvaavien muuttujien havaituista arvoista.
- Tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina *epävarmuutta* ja *satunnaisuutta*.
- Tilastollisesta aineistosta (otoksesta) voidaan laskea erilaisia tunnuslukuja todennäköisyyslaskennan keinoin.

Esimerkki 1/2

Tutkittaessa eräiden komponettien kestoikää valittiin umpimähkään 36 komponettia, joiden kestoikä määritettiin ajannopeutuskokeen avulla. Kestoiät täysinä tunteina:

86	5998	450	6672	988	1014	73	1448	4607
119	506	1113	1901	219	910	34	221	387
793	3036	546	2875	1350	104	3476	2082	2708
2171	4988	2750	3198	427	250	2467	924	40

- Saadaan otoskeskiarvo $\bar{x} = 1693$h, mediaani $Md = (998 + 1014)/2 = 1001$h, otoskeskihajonta $s = 1754$ h, otoskeskipoikkeama $t = \frac{1}{n}\sum_{k=1}^{n} |x_k - \bar{x}| = 292$ h, ensimmäinen kvartiili $Q_1 = 250$ h, kolmas kvartiili $Q_3 = 2708$ h ja kvartiilipoikkeama $Q_3 - Q_1 = 2458$ h.

Esimerkki 2/2

- Luokitellaan aineisto. Valitaan jaoksi 0-999,1000-1999,... ,6000-6999. Merkitään luokkia Lk, $k = 1, \ldots, 7$
- Luokkien frekvensseiksi saadaan:

Luokka	L1	L2	L3	L4	L5	L6	L7
Frekvenssi	18	5	6	3	2	1	1
Kumulat. frekv.	18	23	29	32	34	35	36

- Luokitellusta aineistosta voidaan laskea likiarvo keskiarvolle

$$\bar{x}_l = \frac{1}{n} \sum_{k=1}^{l} f_k \cdot z_k = 1749.5,$$

missä l on luokkien lukumäärä, f_k:t ovat luokkafrekvenssit ja z_k luokkien keskikohdat (esim. tapauksessa 499.5, 1499.5, jne.). Vrt. luokittelemattomasta aineistosta laskettuun keskiarvoon 1693.

Tilastollinen malli

- *Tilastollisella mallilla* tarkoitetaan tutkimuksen kohteita kuvaavien satunnaismuuttujien todennäköisyysjakaumaa, jonka ajatellaan generoineen ko. satunnaismuuttujien havaitut arvot.
- Nämä todennäköisyysjakaumat riippuvat tavallisesti *parametreista*, joiden arvoja ei yleensä tunneta.
- Tilastollista mallia sovellettaessa kohdataan tavallisesti seuraavat *parametreja* koskevat ongelmat:
  - Parametrien arvoja ei tunneta ja ne on *estimoitava* eli arvioitava havaintoaineistosta.
  - Parametrien arvoista on olemassa oletuksia, joita halutaan *testata* havaintoaineiston antaman informaation avulla.

Yksinkertainen satunnaisotos

- Olkoot

$$X_1, X_2, \ldots, X_n$$

riippumattomia, *identtisesti jakautuneita* satunnaismuuttujia, joilla on *sama* pistetodennäköisyys- tai tiheysfunktio $f(x)$.

- Tällöin satunnaismuuttujat

$$X_1, X_2, \ldots, X_n$$

muodostavat *yksinkertaisen satunnaisotoksen* jakaumasta $f(x)$.

Otostunnusluvut ja otosjakaumat

- Olkoon
$$T = g(X_1, X_2, \ldots, X_n)$$
jokin satunnaismuuttujien $X_1, X_2, \ldots, X_n$ (mitallinen) funktio.
- Satunnaismuuttujaa T kutsutaan *otostunnusluvuksi*.
- Tunnusluvun T jakaumaa kutsutaan T:n *otosjakaumaksi*.

Aritmeettisen keskiarvon odotusarvo ja varianssi

- Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ muodostavat yksinkertaisen satunnaisotoksen satunnaismuuttujan X jakaumasta, jonka odotusarvo ja varianssi ovat $E(X) = \mu$ ja $Var(X) = \sigma^2$.
- Havaintojen aritmeettisen keskiarvon $\bar{X}$ odotusarvo ja varianssi ovat

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

- Aritmeettisen keskiarvon standardipoikkeamaa $D(\bar{X}) = \sigma/\sqrt{n}$ kutsutaan *keskiarvon keskivirheeksi*.

Esimerkki (Laininen) 1/2

- Laudan paksuuden X keskihajonta on 0.05 tuumaa ja odotusarvo tuntematon μ tuumaa.
- Kuinka monen umpimähkään valitun laudan paksuus tulee mitata, jotta mitattu keskiarvo poikkeaisi tarkasta odotusarvosta korkeintaan 0.01 tuumaa todennäköisyydellä 0.95?
- Merkitään $\overline{X}$:llä n:n havainnon aritmeettista keskiarvoa. Sen odotusarvo on μ ja varianssi $0.05^2/n$.
- Keskeisen raja-arvolauseen mukaan

$$\mathrm{X} \sim N(\mu, 0.05^2/n).$$

Esimerkki (Laininen) 2/2

- Etsitään n siten, että $Pr(|\overline{X} - \mu| \leq 0.01) = 0.95$, eli

$$Pr\left(|Z| \leq \frac{0.01}{0.05/\sqrt{n}}\right) = 0.95.$$

- Taulukosta saadaan $Pr(|Z| \leq 1.96) = 0.95$, joten ehto on täytetty, jos

$$\frac{0.01}{0.05/\sqrt{n}} = 1.960.$$

- Saadaan $n = 96.04$, eli pyöristettynä ylöspäin $n = 97$.

Aritmeettisen keskiarvon otosjakauma, kun otos on normaalijakautunut

- Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ muodostavat yksinkertaisen satunnaisotoksen normaalijakaumasta $N(\mu, \sigma^2)$.
- Tällöin havaintojen aritmeettinen keskiarvo $\bar{X}$ *noudattaa eksaktisti normaalijakaumaa*:

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

- Standardoitu satunnaismuuttuja

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

noudattaa eksaktisti standardoitua normaalijakaumaa:

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

Otosvarianssin odotusarvo ja varianssi

- Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ muodostavat yksinkertaisen satunnaisotoksen satunnaismuuttujan X jakaumasta, jonka odotusarvo ja varianssi ovat $E(X) = \mu$ ja $Var(X) = \sigma^2$.
- Havaintojen otosvarianssin s^2 odotusarvo ja varianssi ovat

$$E(s^2) = \sigma^2$$

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

Otosvarianssin otosjakauma, kun otos on normaalijakautunut

- Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ muodostavat yksinkertaisen satunnaisotoksen normaalijakaumasta $N(\mu, \sigma^2)$.
- Tällöin satunnaismuuttuja

$$V = \frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

noudattaa eksaktisti χ^2*-jakaumaa* vapausastein $(n-1)$:

$$V \sim \chi^2(n-1)$$

Vinous

- Olkoot

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

  havaintoarvojen 2. ja 3. *keskusmomentti*.

- Tunnuslukua

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

  käytetään kuvaamaan havaintoarvojen jakauman *vinoutta*.

  - Jos $c_1 \approx 0$, on havaintoarvojen jakauma *symmetrinen painopisteensä suhteen*.
  - Jos $c_1 > 0$, on havaintoarvojen jakauma *positiivisesti vino*.
  - Jos $c_1 < 0$, on havaintoarvojen jakauma *negatiivisesti vino*.

Huipukkuus

- Olkoot

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

  havaintoarvojen 2. ja 4. *keskusmomentti*.

- Tunnuslukua

$$c_2 = \frac{m_4}{m_2^2} - 3$$

  käytetään kuvaamaan havaintoarvojen jakauman *huipukkuutta*.
  - *Normaalijakautuneella* havaintoaineistolla $c_2 \approx 0$.
  - Jos $c_2 > 0$, on havaintoarvojen jakauma *huipukas* (Normaalijakautuneeseen havaintoaineistoon verrattuna).
  - Jos $c_2 < 0$, on havaintoarvojen jakauma *laakea* (Normaalijakautuneeseen havaintoaineistoon verrattuna).

Geometrinen- ja harmoninen keskiarvo

- Aritmeettinen keskiarvo ei ole kaikissa tilanteissa sopiva keskiluku. Vaihtoehtoisia keskilukuja ovat *geometrinen keskiarvo* ja *harmoninen keskiarvo*.
- Havaintoarvojen $x_1, x_2, \ldots, x_n$ *geometrinen keskiarvo* on

$$G = \sqrt[n]{x_1, x_2, \ldots, x_n}.$$

- Havaintoarvojen $x_1, x_2, \ldots, x_n$ *harmoninen keskiarvo* on

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}.$$

Esimerkki

- Rahasto-osuuden arvo kasvaa ensimmäisenä vuotena 5.5%, toisena 9.3%, kolmantena 1.2% ja neljäntenä 7.1%. Mikä on rahaston keskimääräinen vuotuinen tuotto?
- Tässä on luonnollista käyttää geometrista keskiarvoa. Saadaan

$$\sqrt[4]{1.055 \cdot 1.093 \cdot 1.012 \cdot 1.071} \approx 1.05733.$$

- Tämä luku on luonnollinen siinä mielessä, että

$$1.055 \cdot 1.093 \cdot 1.012 \cdot 1.071 \approx 1.2498 \approx 1.05733^4.$$

Esimerkki

- Matka TKK:n matematiikan laitokselta (Otakaari 1, Espoo) Tilastokeskukseen (Työpajankatu 13, Helsinki) julkisilla kulkuvälineillä vaatii:
  - 5 km bussilla 102 (keskinopeus 55km/h) $\approx$ 5.45 min
  - 4 km metrolla (keskinopeus 70km/h) $\approx$ 3.43 min
  - 0.7 km kävely (keskinopeus 5km/h) $\approx$ 8.40 min
- Keskinopeus on siis $60(5+4+0.7)/17.27 \approx 33.7$ km/h.
- Laskemalla harmoninen keskiarvo saadaan

$$1/((5/55+4/70+0.7/5)/9.7) \approx 33.7 \text{ km/h}.$$