Sovellettu todennäköisyyslaskenta B

Antti Rasila

11. lokakuuta 2007

1. Johdantoa tilastotieteeseen
   - Peruskäsitteitä
   - Tilastollisen kuvailun ja päättelyn menetelmiä
   - Kontrolloidut kokeet
   - Tilastolliset mitta-asteikot

2. Aineistojen kuvaaminen
   - Frekvenssit ja havaintoarvojen jakauma
   - Tunnuslukuja suhdeasteikolliselle aineistolle
   - Standardointi ja tilastollinen etäisyys
   - Tunnuslukuja järjestysasteikollisille aineistolle

Mitä tilastotiede on?

- Sana tilasto (statistics) viittaa valtioon. Hallitsijat halusivat saada tietoa asukkaistaan, kaupasta, jne. valtion asioiden hoitamista varten.
- Tilastotiede kehittää ja soveltaa menetelmiä, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä tilanteissa, joissa ilmiöitä koskeviin tietoihin liittyy epävarmuutta ja satunnaisuutta.
- Tilastotiede on todennäköisyyslaskennan tärkeimpiä sovellusalueita.

Peruskäsitteitä

- *Tilastolla* tarkoitetaan johonkin joukkoon liittyvää numeerista informaatiota, joka saadaan yhdistelemällä ko. joukon yksittäisiin alkioihin liittyviä tietoja.
- Tarkastelun kohteena olevaa joukkoa kutsutaan *perusjoukoksi* eli *populaatioksi* ja sen alkioita *yksiköiksi*.
- Tilastoa tarkastellaessa tutkitaan jotakin yksiköiden *ominaisuutta*.
- Luotettavin tulos saadaan, jos tutkitaan jokainen perusjoukon alkio erikseen. Yleensä tämä ei ole mahdollista.
- Siksi yleensä tyydytään tutkimaan perusjoukkoa ottamalla siitä jokin *otos*.

Tilastollisen kuvailun ja päättelyn menetelmiä

- Kuvailun menetelmiä:
  - Tilastografiikka
  - Tilastolliset tunnusluvut
  - Tilastolliset mallit
- Päättelyn menetelmiä:
  - Tilastolliset mallit
  - Tilastollinen testaus

Aineiston kerääminen

- Kohdistuuko tutkimus koko perusjoukkoon vai vain johonkin sen osaan?
  - Tutkimusta kutsutaan kokonaistutkimukseksi, jos perusjoukon kaikki alkiot tutkitaan.
  - Tutkimusta kutsutaan otantatutkimukseksi, jos perusjoukon alkioista vain osa tutkitaan. (Otoksen valitsemiseen on useita erilaisia tapoja.)
- Muutetaanko tutkimuksessa aktiivisesti tutkimuksen kohteiden olosuhteita?
  - Tutkimus on *koe*, jos tutkitaan olosuhteiden muuttamisen vaikutusta tutkimuksen kohteisiin.
  - Jos olosuhteita ei muuteta aktiivisesti, sanomme, että tutkimus perustuu *suoriin havaintoihin*

Kontrolloidut kokeet

- Kokeesta ei voida tehdä luotettavia johtopäätöksiä, ellei koe ole *kontrolloitu*:
  - Kokeessa on *vertailtava* vähintään kahden erilaisen käsittelyn vaikutuksia.
  - Käsittelyjen kohdistamisessa on käytettävä *satunnaistusta*.
  - Kokeessa on tehtävä *riittävästi koetoistoja*.

Mittaaminen

- Tilastollisen tutkimuksen kohdetta kuvaavat numeeriset ja kvantitatiiviset tiedot saadaan mittaamalla.
- Mittari on funktio, joka liittää mitattavan kohteen ominaisuuteen numeerisen arvon.

Tilastolliset mitta-asteikot

- *Nominaali- eli laatueroasteikkoa* voidaan käyttää silloin, kun osataan erottaa keskenään erilaiset yksiköt. Esimerkiksi omena vai appelsiini?
- *Ordinaali- eli järjestysasteikkoa* käytetään silloin, kun tilastoyksiköt voidaan jakaa luokkiin, joiden välillä on järjestys. Esim. oppiarvot: ylioppilas, kandidaatti, di/maisteri, lisensiaatti, tohtori.
- *Intervalli- eli välimatka-asteikko*. Esim. lämpötila Celsius-asteina.
- *Suhdeasteikossa* on lisäksi absoluuttinen nollakohta. Suhdeasteikolla on mielekästä kysyä, kuinka monta kertaa enemmän tai vähemmän. Esimerkiksi pituus tai lämpötila Kelvin-asteikolla.
- Kvalitatiivisia ominaisuuksia kuvataan laatueroasteikollisilla muuttujilla ja kvantitatiivisia puolestaan välimatka- tai suhdeasteikollisilla muuttujilla.

Frekvenssit ja havaintoarvojen jakauma

- Jos muuttuja on diskreetti, havaittujen arvojen jakaumaa kuvataan *frekvessijakaumalla* ja sitä vastaavalla graafisella esityksellä, joka on *pylväsdiagrammi*.
- Jos muuttuja on jatkuva, havaittujen arvojen jakaumaa kuvataan *luokitellulla frekvessijakaumalla* ja sitä vastaavalla graafisella esityksellä, joka on *histogrammi*.
- Histogrammissa pinta-ala vastaa frekvenssiä ja pylväsdiagrammissa korkeus.
- Muuttujan x mahdolliset arvot $y_1, y_2, \ldots, y_m$ yhdessä niiden havaittujen frekvenssien $f_1, f_2, \ldots, f_m$ kanssa muodostavat muuttujan x havaittujen arvojen $x_1, x_2, \ldots x_n$ frekvenssijakauman.
- Siinä missä todennäköisyyksien summa on aina yksi, on nyt $\sum f_i = n$.

Tunnuslukuja suhdeasteikolliselle aineistolle 1/2

- Aritmeettinen keskiarvo:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Otosvarianssi:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

- Otoskeskihajonta: $s = \sqrt{s^2}$

Tunnuslukuja suhdeasteikolliselle aineistolle 2/2

- Origomomentit:

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

- Keskusmomentit:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Standardointi ja tilastollinen etäisyys

- Standardoitujen havaintoarvojen

$$z_i = \frac{x_i - \bar{x}}{s_x}, i = 1, 2, \ldots, n$$

aritmeettinen keskiarvo ja otosvarianssi ovat

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = 0 \qquad s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = 1$$

- Havaintoarvojen x_k ja x_l tilastollinen etäisyys d_{kl} on

$$d_{kl} = \frac{x_k - x_l}{s_x}$$

Tunnuslukuja järjestysasteikollisille aineistolle

- Järjestystunnusluvut: Suuruusjärjestyksessä k. havaintoarvoa z_k kutsutaan k. järjestystunnusluvuksi.
- *Minimi* ja *maksimi* eli pienin ja suurin arvo.
- *Vaihteluväli* ja sen pituus.
- *Prosenttipisteet* z_p: p. prosenttipiste jakaa aineiston kahteen osaan: p% havainnoista on prosenttipistettä pienempiä ja loput $(100 - p)$% suurempia.
- *Mediaani* eli $Me = z_{50}$ jakaa aineiston kahteen yhtä suureen osaan.
- *Kvartiilit*: $Q_1 = z_{25}$, $Q_2 = z_{50} = Me$ ja $Q_3 = z_{75}$.
- *Kvartiilipoikkeama*: $(Q_3 - Q_1)/2$

Tunnuslukuja laatueroasteikollisille aineistoille

- *Suhteelliset frekvenssit* f_i/n.
- *Moodi eli tyyppiarvo* eli yleisin havaintoarvo.