Data-analyysi R-ohjelmistolla

Tommi Härkänen

Terveyden ja hyvinvoinnin laitos (THL), Helsinki
E-mail: tommi.harkanen@helsinki.fi

Helsingin yliopisto, 1.4.2014

Sisältö

Koe

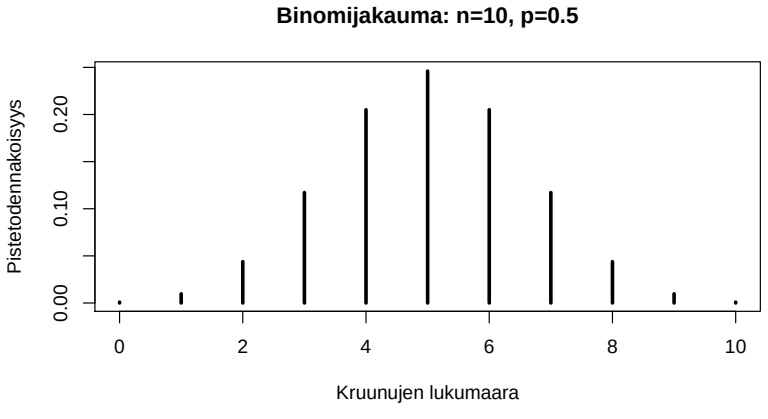
- Tutkijalla on hypoteesi (oletus).
- Tutkija suunnittelee ja toteuttaa kokeen sekä kerää havaintoaineiston.
- **Kysymys** Tukeeko havaintoaineisto tutkijan hypoteesia?
- Esim. lantinheitto:
  - Hypoteesi "Kruunan todennäköisyys on 0.5".
  - Koe "Heitetään lanttia n kertaa".
  - Havaintoaineisto Kruunujen lukumäärä n heitossa.
- Esim. lääketieteellinen koe:
  - Hypoteesi Hoidolla ei ole tehoa.
  - Koe "Jaetaan potilaat satunnaisesti hoito- ja kontrolliryhmiin. Annetaan hoitoa hoitoryhmässä ja lumehoitoa kontrolliryhmässä".
  - Havaintoaineisto Potilaiden paranemishavainnot ja tieto ryhmästä.

Hypoteettisten toistokokeiden jakauma

Esim. binomikoe.

Tutkijan hypoteesi Kruunan ($Y_i = 1$) todennäköisyys lantinheitossa on $p = 0.5$.

Havaintoaineisto Kruunujen lukumäärä n heitossa $\sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, p)$.

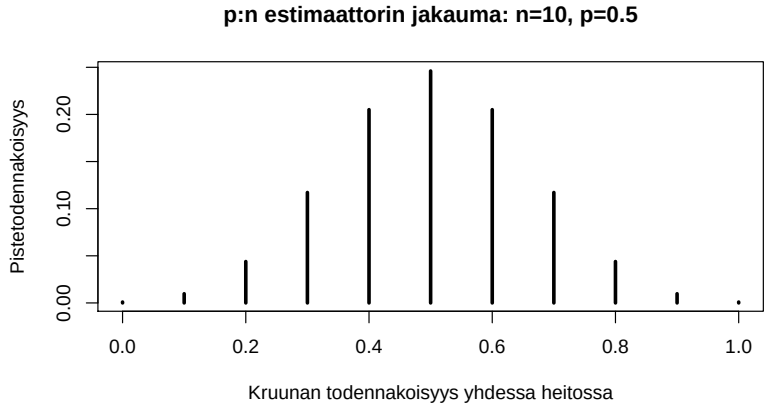


Otantajakauma

- Frekventistisessä päättelyssä oletetaan, että koe voidaan toistaa (rajattoman monta kertaa) samanlaisissa olosuhteissa.
- Tutkija on kiinnittänyt todennäköisyysmallin $f(\mathbf{y}; \theta)$ ja siihen liittyvien parametrien θ arvot asettamansa hypoteesin mukaisesti.
- Tutkija valitsee tunnusluvun $t(\mathbf{Y})$, esim. estimaattorin jakauman odotusarvolle (keskiarvo).
- Tutkija muodostaa otantajakauman tunnusluvulle $t(\mathbf{Y})$.
- Tutkija laskee havaintoaineistosta $\mathbf{y}$ tunnusluvun arvon $t(\mathbf{y})$ ja vertaa tätä hypoteesin mukaiseen otantajakaumaan.

Tunnusluvun otantajakauma binomikokeessa

Todennäköisyyttä $\mathbb{P}\{Y_i = 1\} = p$ voidaan estimoida ykkösten suhteellisella osuudella havaintoaineistossa $\hat{p}(\mathbf{Y}) = \sum_{i=1}^{n} Y_i/n$. Estimaattorin todennäköisyysjakauma – kun n ja p on kiinnitetty (tutkijan hypoteesin mukaisesti) – on



Normaalijakautuneiden satunnaismuuttujien keskiarvo

SD on tunnettu

Oletetaan n riippumatonta satunnaismuuttujaa $Y_i \sim \mathrm{N}(\mu, \sigma^2)$. Parametri μ on odotusarvo ja σ^2 varianssi (σ on keskihajonta SD).

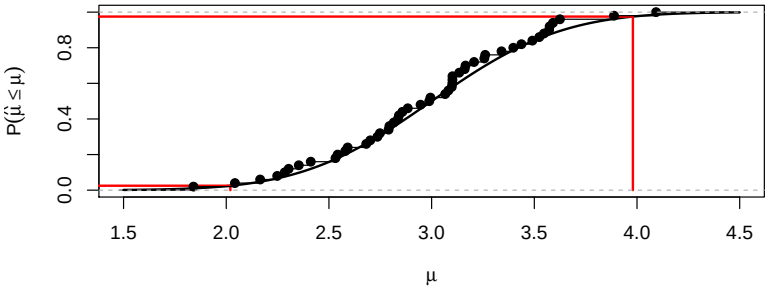
Satunnaismuuttujien keskiarvo (parametrin μ estimaattori) on myös normaalijakautunut $\hat{\mu} := \sum_{i=1}^{n} Y_i/n \sim \mathrm{N}(\mu, \sigma^2/n)$. Keskiarvon **keskivirhe** SE on $\sigma/\sqrt{n}$.
Suuri $n \Rightarrow$ pieni SE.

Odotusarvoparametrin μ luottamusväli

Luottamusjoukko $A(\mathbf{Y}) \subset \Theta$ määriteltiin $\mathbb{P}\{\mu \in A(\mathbf{Y})\} \geq 1 - \alpha$, jossa *luottamustaso* $1 - \alpha$ on usein 0.95.
Normaalijakaumamallin odotusarvon tapauksessa luottamusväli määritellään usein symmetrisenä

$$\hat{\mu} - \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n} < \mu < \hat{\mu} + \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n},$$

jossa standardinormaalijakauman $\mathrm{N}(0, 1)$ kvantiilifunktio on $\Phi^{-1}(\cdot)$, `qnorm()`.



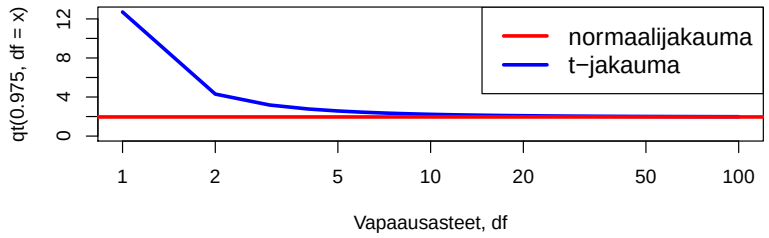
Studentin t-jakauma

SD on tuntematon

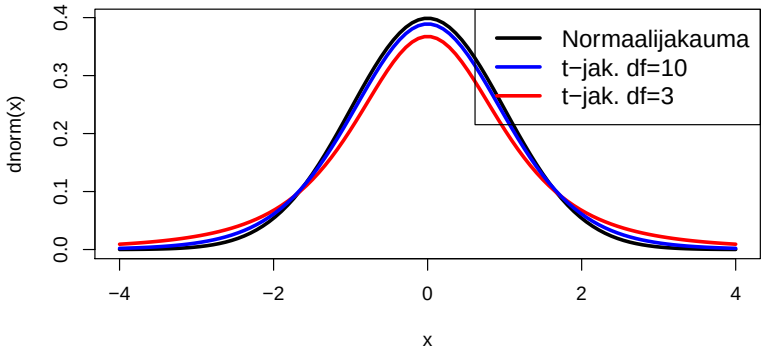
Yleensä σ on tuntematon kuten μ. Jos otoskoko n on suuri, niin havaittu SD s on lähellä σ.

Jos n on pieni, epävarmuus keskihajonnassa s pitää huomioida käyttämällä t-jakaumaa $n-1$ vapausasteella (df) normaalijakauman sijaan.

97.5% kvantiilit t-jakaumalla verrattuna normaalijakaumaan:



t-jakauma vs. normaalijakauma



Jos havaintoja n on vähän (df on pieni), s on epätarkka ja keskiarvon otantajakauma tuottaa enemmän poikkeavain pieniä ja suuria arvoja ("paksut hännät") kuin normaalijakauma.