Data-analyysi R-ohjelmistolla

Tommi Härkänen

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

Helsingin yliopisto, 25.3.2014

Sisältö

Mitä on satunnaisuus?

Soveltavissa tieteissä ennustettavuuden puute:

Deterministinen tulos Esim. kaikki hoidon saaneet potilaat paranevat, mutta ilman hoitoa jääneet eivät parane.

Satunnainen tulos Esim. hoidetuista 60 % paranee, mutta ilman hoitoa jääneistä vain 20 %.

Muita esimerkkejä satunnaisuudesta

- Kolikon tai nopan heitto
- Kvanttimekaniikka
- Sää
- Osakemarkkinat

Todennäköisyyden määritelmiä

Yksittäistä tapahtumaa ei voi ennustaa, jos se ei ole varma tapahtuma. Voidaan muodostaa **subjektiivisia todennäköisyyksiä** eri vaihtoehdoille ennen tapahtuman havaitsemista.

Jos prosessi, joka tuottaa havaintoaineiston, voidaan **toistaa**, eri vaihtoehtojen tapausmäärät ja suhteelliset osuudet voidaan laskea $\Rightarrow$ **frekvenssitodennäköisyys**.
(Ongelma: voidaanko olla varmoja, että olosuhteet eri toistokerroilla säilyvät samoina?)

Jatkuvat vs. diskreetit todennäköisyysjakaumat

- **Diskreetti** satunnaismuuttuja voi saada äärellisen määrän eri arvoja tai arvot voidaan numeroida 1, 2, 3, ... Eri vaihtoehtojen todennäköisyyksien summa on 1, ja yksittäinen arvo voi saada positiivisen todennäköisyyden.
  - Nopan heitto (6 mahdollista arvoa)
  - Klaavojen lukumäärä ennen ensimmäistä kruunaa lantinheitossa (äärellinen määrä mahdollisia arvoja 0, 1, 2, ...)
- **Jatkuvan** satunnaismuuttujan todennäköisyys saada jokin yksittäinen arvo on **nolla**. Esim.
  - Ihmisen pituus.
  - Verenpaine.

Todennäköisyysjakauman kuvaaminen

Kertymäfunktio Todennäköisyys että satunnaismuuttujan X arvo **korkeintaan** x: $\mathbb{P}\{X \leq x\}$. Esim. todennäköisyys, että satunnaisesti valitun opiskelijan pituus on korkeintaan $x = 170$ cm.

Tiheysfunktio (Piste-) todennäköisyys että **diskreetti** satunnaismuuttuja X saa arvon x voi olla nolla tai positiivinen. Esim. todennäköisyys että satunnaisesti valittu opiskelija on mies $\mathbb{P}\{X = \text{mies}\}$.
Jatkuvalle satunnaismuuttujalle tiheysfunktion arvo pisteessä x on $f(x) \geq 0$. Tiheysfunktion arvo kerrottuna pienellä vakiolla $\epsilon > 0$ on likimain todennäköisyys että satunnaismuuttujan arvo on (lyhyellä) välillä $[x, x + \epsilon]$.

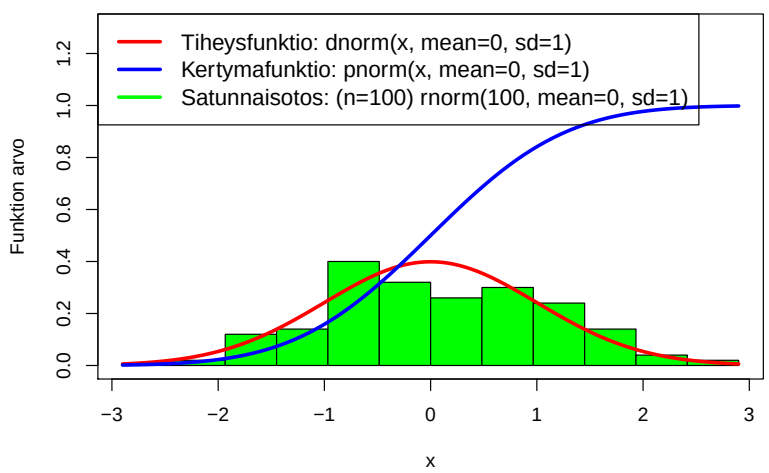
Jakaumat R-ohjelmistossa

- Useimmista todennäköisyysjakaumista R:stä löytyy 4 funktiota, jotka eroavat alkukirjaimen perusteella:
  - d tiheysfunktio
  - p kertymäfunktio
  - q kvantiilifunktio (kertymäfunktion käänteisfunktio)
  - r satunnaislukujen generointi
- Esim. normaalijakaumalle löytyvä funktiot

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

- R:n perusasennuksen mukana tulevat jakaumat on lueteltu manuaalisivulla `Distributions` ja lisää löytyy CRAN-sivustolta, Task Views -sivun kohdasta Distributions.

Normaalijakauma



Havaintoaineisto, malli ja parametrit

Edellä poimittiin otos $\mathbf{y} = (y_i)$, $i = 1, 2, \ldots, n$ satunnaismuuttujien Y_i realisaationa. Jakauman määräävien parametrien θ arvot olivat tunnettuja. Jos satunnaismuuttujat Y_i ovat riippumattomia (ehdolla θ), niin jakauman yhteistiheys- tai yhteispistetodennäköisyysfunktio voitiin kirjoittaa tulona

$$L(\mathbf{y};\, \theta) = f(\mathbf{y};\, \theta) = \prod_{i=1}^{n} f(y_i;\, \theta).$$

Tässä siis θ on kiinteä, vakio, ja $\mathbf{y}$ vapaita parametreja.

Empiirisessä tutkimuksessa parametrien θ arvot ovat tuntemattomia, ja kokeen avulla kerätyn havaintoaineiston $\mathbf{y}$ avulla pyritään estimoimaan tuntemattomien parametrien θ arvot.

Tutkija usein valitsee sopivan malliperheen, jonka indeksoi $\theta \in \Theta$, ja josta valitaan parhaiten havaintoaineistoon sopiva malli.

Uskottavuusfunktio

Havaintoaineiston keräämisen jälkeen $\mathbf{y}$ ajatellaan kiinteäksi, ja uskottavuusfunktio määritellään **parametrien θ funktiona** $L(\mathbf{y};\, \theta) = f(\mathbf{y};\, \theta) = \prod_{i=1}^{n} f(y_i;\, \theta)$.
Esimerkki: Havaintoja normaalijakaumasta, jonka keskihajonta on 1. Tuntematon θ on odotusarvoparametri.

Suurimman uskottavuuden (SU) estimointi

Parametrien θ arvon, joka maksimoi uskottavuusfunktion arvon, ajatellaan olevan uskottavin (tuntemattoman) parametrin arvo annetun aineiston perusteella.

R-ohjelmistosta löytyy useita funktioita, joiden avulla voidaan hakea annetun funktion minimi- tai maksimikohtaa. SU-estimoinnin kannalta `stats4`-paketista löytyvä funktio `mle()` (joka minimoi annetun funktion) on monipuolinen. Tärkeimpiä argumentteja ovat:

- minuslogl Negatiivinen, logaritminen uskottavuusfunktio $-\ell(\theta)$ (**!**).
- start Alkuarvot minimointialgoritmille

Huomattavaa mle()-funktion käyttämisessä: Jos parametriavaruus Θ on rajoitettu, niin uskottavuusfunktion pitää palauttaa puuttuva arvo (`NA`), jos $\theta \notin \Theta$.