Satunnaistamistestaus

Satunnaistamistestaus: sisältö

- satunnaistamisen periaate
- p-arvon estimointi
- kahden joukon vertaaminen
- parittainen testaus
- riippumattomuuden testaaminen
- Lähde: Cohen, luvut 4.2, 4.3, 5.3

Hypoteesin testaus: kertausta

- hypoteesi
- nollahypoteesi
- p-arvo: todennäköisyys saada havaittu tulos nollahypoteesin vallitessa
- pieni p-arvo → tilastollisesti merkitsevä tulos
- miten valitaan nollahypoteesi?
- miten estimoidaan nollahypoteesin mukaista tunnusluvun jakaumaa?

Nollahypoteesi

Valinta, esim.

- kahden joukon tai otoksen vertaaminen
  - nollahypoteesi: joukot ovat samasta jakaumasta
- riippuvuus tai korrelaatio asioiden välillä
  - nollahypoteesi: yhteydet ovat satunnaisia

Testaaminen

- sekoitetaan data nollahypoteesin mukaisesti
  = permutation test = (re)randomization test
- generoidaan monta permutointia

→ saadaan nollahypoteesin mukainen jakauma

Esimerkki

Sovellus: opiskelijavalinta

- valintaperusteita muutettiin
- käytössä myös erillinen, samana pysynyt testi valituille opiskelijoille
- vaihtelu testin pisteissä näyttää kasvaneen
- mittana kvartaalivälin pituus: IQR = 0.75-fraktiili – 0.25-fraktiili
- testattava muuttuiko tulos merkittävästi
  - käytettävissä 25 opiskelijan tulokset viime vuodelta
  - 20 opiskelijan tulokset tältä vuodelta

Esimerkki

- viime vuonna IQR = 6, tänä vuonna IQR = 8.5
- erotus $d = 6 - 8.5 = -2.5$
- millä todennäköisyydellä erotus $d = -2.5$ saatiin sattumalta?
- nollahypoteesi: vaihtelu ei tänä vuonna ole viime vuotta suurempaa, vaan opiskelijat ovat samasta (tuntemattomasta) jakaumasta

→ kuka tahansa opiskelijoista voisi olla kummassa ryhmässä tahansa

→ nollahypoteesin mukainen tilanne saadaan sijoittamalla opiskelijat ryhmiin satunnaisesti

Kahden joukon vertaaminen

- **Algoritmi:**
- Olkoon S_A ja S_B kaksi otosta, joiden koot ovat N_A ja N_B. Olkoon $\theta = f(S_A, S_B)$ otoksista laskettu tunnusluku.
- Toista K kertaa:
  - sekoita joukkoyhdisteen $S_{A+B} = S_A \cup S_B$ alkiot
  - sijoita N_A ensimmäistä pseudo-otokseen A_i^* ja loput N_B alkiota B_i^*:hin
  - laske $\theta_i^* = f(A_i^*, B_i^*)$
- Käytä θ_i^*:n jakaumaa sen testaamiseen, mikä on θ:n todennäköisyys nollahypoteesin vallitessa

p-arvon arvioiminen

- oletetaan, että vaihtoehtoinen hypoteesi on muotoa *θ on suurempi kuin voisi odottaa sattumalta*
- $p = P(\theta_i^* \geq \theta \mid \text{nollahypoteesi on voimassa}) \approx \frac{|\{\theta_i^* \geq \theta | 1 \leq i \leq K\}|}{K}$
- vastaavasti toisin päin jos "$\leq$" eikä "$\geq$"
  - nämä ovat "yksihäntäisiä" testejä
- jos hypoteesi on muotoa θ *poikkeaa sattumanvaraisesta*, käytetään kaksihäntäistä testiä:
  - jos $\theta \geq \theta_i^*$:n 0.5-fraktiili: $p \approx 2\frac{|\{\theta_i^* \geq \theta | 1 \leq i \leq K\}|}{K}$
  - jos $\theta \leq \theta_i^*$:n 0.5-fraktiili: $p \approx 2\frac{|\{\theta_i^* \leq \theta | 1 \leq i \leq K\}|}{K}$

Parittainen testi

- kaksi robottia osallistuu robottien 10-otteluun
- A saa tehtävistä keskimäärin enemmän pisteitä kuin B
- onko A merkitsevästi parempi kuin B?
- nollahypoteesi: robotit ovat yhtä hyviä

→ kunkin ottelun tulos olisi voinut yhtä hyvin olla toisin päin

- satunnaistaminen: vaihda satunnaisesti tehtävien sisällä robottien tuloksia keskenään
- huom: kukin pistemäärä liittyy tiettyyn tehtävään

Riippumattomuustesti

- tarkastellaan esim. kahta aineiston attribuuttia
- onko niillä merkitsevää tilastollista riippuvuutta?
- **Algoritmi:**
- Olkoot S_A muuttujan A saamien arvojen jono ja S_B muuttujan B saamien arvojen jono. Olkoon $\theta = f(S_A, S_B)$ jonoista laskettu tunnusluku (esim. korrelaatio)
- Toista K kertaa:
  - sekoita S_B:n järjestys satunnaiseksi
  - laske $\theta_i^* = f(S_A, S_B)$
- Käytä θ_i^*:n jakaumaa sen testaamiseen, mikä on θ:n todennäköisyys nollahypoteesin vallitessa

Satunnaistamistesti

- ei tarvitse oletuksia jakaumista ja niiden parametreista
- ei estimoi populaation parametreja
- käyttää vain saatavilla olevaa dataa
- tarvitsee kaksi aineistoa tai kaksi muuttujaa, joita vertaillaan
- tuottaa likimääräisen p-arvon
  - täydellinen permutointi tuottaisi tarkan arvon, mutta harvoin mahdollinen
- voi olla laskennallisesti vaativaa

Esimerkki

- sovellus: geenikartoitus eli tietylle perinnölliselle sairaudelle altistavan geenin paikantaminen
- aineisto koostuu sairaista ja terveistä ihmisistä
- lisäksi tietoja kunkin ihmisen perimän eri kohtien ominaisuuksista
- kartoitusmenetelmän ensimmäinen vaihe muodostaa kullekin perimän kohdalle mallin, joka ennustaa onko henkilö sairas vai terve
  - taudille altistavan geenin kohdalle tehty malli erottelee sairaat hyvin terveistä
  - mutta hyvä malli voi löytyä sattumalta

Esimerkki

- mallin hyvyyden arviointi satunnaistamistestillä
  - riippumattomuustesti: sairas/terve-status ei riipu perimästä
  - satunnaistetaan statukset
  - kuinka hyviä malleja menetelmä onnistuu löytämään nollahypoteesin vallitessa?
- geenin ennustetaan olevan sillä kohdalla, joka saa parhaan p-arvon

Yhteenveto

- mitä testattava hypoteesi koskee?
  - usein jokin "rakenteellinen" asia: kahden joukon suhde, kahden muuttujan suhde, asioiden järjestykseen liittyvä asia
- nollahypoteesia vastaa yleensä rakenteen satunnainen sekoittaminen
- toista satunnainen sekoittaminen monta kertaa
  - p-arvo on niitten satunnaistamisten osuus, jotka ovat vähintään yhtä "hyviä"
  - saatu p-arvo on likimääräinen
  - karkea nyrkkisääntö: kohtuullinen tarkkuus $K = 50/p$ satunnaistamisella