Tilastollisia peruskäsitteitä ja Monte Carlo

Tilastollisia peruskäsitteitä

- hypoteestin testaus
- merkitsevyystaso, p-arvo
- otosjakauma
- otosjakauman Monte Carlo -estimointi

Esimerkki

- kaksi shakkiohjelmaa, A ja B, pelaavat 15 ottelua
- A voittaa niistä 10 ja B voittaa 5
- A voitti siis näistä otteluista osuuden $f = 0.67$
  - f on aineistosta laskettu tunnusluku
- mutta mikä on A:n todennäköisyys π voittaa uusi peli B:tä vastaan?
  - π on ilmiön "todellinen" tunnusluku (kaikille mahdollisille peleille)
- voi olla mahdotonta saada selville oikeaa π:n arvoa
- onko A varmasti parempi kuin B? Kuinka varmasti?

Hypoteesin testaus

- onko toinen shakkiohjelmista toista parempi?
- jos ohjelmat ovat yhtä hyviä, niin $\pi = 0.5$
  - = nollahypoteesi
- (vaihtoehtoinen) hypoteesi: A on parempi kuin B
- jos nollahypoteesi on voimassa, kuinka todennäköistä on havaita 15 ottelussa tulos $f = 0.67$
- jos epätodennäköistä, hylätään nollahypoteesi $\pi = 0.5$
- hypoteesi → nollahypoteesi → nollahypoteesin testaus → nollahypoteesin hylkääminen?

Satunnaismuuttuja

- satunnaismuuttuja X on suure, joka saa eri arvoja eri todennäköisyyksillä
  - esim. $X = 1$ jos A voitti pelin ja $X = 0$ jos A hävisi
- satunnaismuuttujalla voi olla tietty hyvin määritelty arvo $X = x$
  - X on muuttuja, x on sen arvo
  - kun peli on pelattu, X:llä on arvo 0 tai 1
- satunnaismuuttujalla ja sen jakaumalla on erilaisia tunnuslukuja, esim.
  - X:n odotusarvo EX
  - X:n varianssi σ_X

Tunnuslukujen estimointi

- käytettävissä oleva aineisto on otos ilmiöstä
- $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$
  - otoskeskiarvo $\bar{x}$, otosvarianssi σ_x
  - todellinen ilmiön keskiarvo EX, todellinen varianssi σ_X
- kuinka hyviä arvioita otoksesta lasketut tunnusluvut ovat jakauman tunnusluvuille?

Iid-oletus

- iid-oletus (independent and identically distributed):
  - X_i:tten arvot ovat toisistaan riippumattomia
  - kaikki X_i:t noudattavat samaa jakaumaa
- iid-oletus joudutaan tekemään usein
  - mutta onko se aina voimassa?
- jos iid-oletus on voimassa, niin erityisesti
  - $E(\bar{x}) = EX$
  - $E(\sigma_x) = \sigma_X$
  - otoskeskiarvo ja -varianssi ovat harhattomia estimaatteja

Otosjakauma

- mielivaltainen tunnusluku t
- t:n otosjakauma = tunnusluvulle t tietyn kokoisissa otoksissa (iid) havaittava jakauma
- esim. $t = f =$ se osuus peleistä, jotka A voittaa
  - pelataan 15 peliä, lasketaan f
  - toiset 15 peliä $\rightarrow$ eri arvo f:lle?
- otostunnusluku on satunnaismuuttuja, jonka jakauma on otosjakauma
  - esimerkissä f on satunnaismuuttuja
  - f:n jakauma kaikissa mahdollisissa 15 pelin otteluissa on sen otosjakauma

Otosjakauma ja hypoteesin testaus

- jos shakkiohjelmat A ja B ovat yhtä hyviä niin $\pi = 0.5$
  - tämä on *nollahypoteesi*
- silloin f noudattaa jotain tiettyä otosjakaumaa
  - (binomijakauma, mutta sillä ei tässä ole merkitystä)
- nollahypoteesin mukainen otosjakauma määrää, millä todennäköisyydellä havaitaan $f \geq 0.67$
- jos todennäköisyys $P(f \geq 0.67)$ on suuri, niin havaittu A:n kannalta suotuisa tulos voidaan saada sattumalta

$\Rightarrow$ ei voida päätellä luotettavasti että A on parempi kuin B

- jos todennäköisyys $P(f \geq 0.67)$ on pieni, on epätodennäköistä että A:n kannalta niin suotuisa tulos on saatu sattumalta

$\Rightarrow$ *hylätään* nollahypoteesi $\pi = 0.5$

- $p = P(f \geq 0.67)$ on tuloksen p-*arvo*
  - mitä pienempi p-arvo, sitä tilastollisesti merkitsevämpi tulos
- jos $p \leq 0.05$, niin A on parempi kuin B *merkitsevyystasolla* 0.05
  - tyypillisiä merkitsevyystasoja (raja-arvoja merkitsevyydelle): 0.05, 0.01, 0.001

Monte Carlo -menetelmä

- nollahypoteesin mukainen otosjakauma on usein tuntematon
  - p-arvon laskeminen analyyttisesti voi olla vaikeaa tai mahdotonta
- otosjakaumaa voidaan estimoida helposti
  - simuloi 15 ottelun turnauksia
    - arvo A voittajaksi nollahypoteesin mukaisesti tn:llä 0.5
    - laske tunnusluku f kullekin turnaukselle
  - f:n arvot ovat otos otosjakaumasta

$\Rightarrow$ arvojen jakauma on estimaatti otosjakaumalle

- simuloi esim. $N = 1000$ turnausta, kussakin 15 ottelua
  - arvo A ottelun voittajaksi todennäköisyydellä 0.5
  - laske kussakin turnauksessa f, A:n voittojen osuus
- laske monessako turnauksessa $f \geq 0.67$, merkitään lukumäärää N_0:lla
- $p \approx N_0/N$
- Monte Carlo -algoritmit ovat yleisesti satunnaisuutta hyväksi käyttäviä menetelmiä
  - tilastolliset ongelmat vain yksi sovellusalue
  - nimi tulee kuuluisasta kasinokaupungista

Yhteenveto

- hypoteestin testaus: nollahypoteesi, vaihtoehtoinen hypoteesi
- pieni p-arvo = tilastollisesti merkitsevä tulos
- p-arvo perustuu otosjakaumaan
- otosjakaumaa voidaan estimoida Monte Carlo -menetelmällä
  - esimerkin tilanteessa analyyttinenkin ratkaisu olisi ollut helppo
  - MC-menetelmää voidaan käyttää hyvin vaikeittenkin otosjakaumien estimointiin