Satunnaislukujen generointi

Lähteet

- Knuth, D., The Art of Computer Programming, Volume 2: Seminumerical Algorithms, luku 3
- Numerical Recipes in C, luku 7

Johdanto

- edellä olemme tutustuneet menetelmiin, jotka perustuvat satunnaislukujen generointiin
  - Monte Carlo -testaus: generoidaan arvoja tunnetusta populaatiojakaumasta
  - satunnaistustestaus: muodostetaan (kahden) otoksen permutaatioita
  - bootstrap: generoidaan pseudo-otoksia valitsemalla satunnaisesti alkioita otoksesta takaisinpanolla

Johdanto (2)

- satunnaistus ja bootstrap edellyttävät satunnaislukujen generoimista tasaisesta jakaumasta
  - tasaisesti jakautuneiden satunnaislukujen generointiin ohjelmointikielissä kirjastofunktiot
- Monte Carlo -testauksen kohdalla populaatiojakauma voi luonnollisesti olla mikä tahansa jakauma
  - kuinka generoidaan satunnaislukuja muista kuin tasaisista jakaumista?
  - vaikka populaatiojakauma tunnettaisiinkin, ei välttämättä ole aivan suoraviivaista generoida siitä otoksia

Diskreettejä jakaumia

- satunnaismuuttuja noudattaa aina jotakin todennäköisyysjakaumaa
- jos muuttujan arvot ovat diskreettejä jakauma on diskreetti
- diskreetin todennäköisyysjakauman määrittää pistetodennäköisyysfunktio $p(x)$, jolle $\sum_i p(i) = 1$
- keskeisiä diskreettejä jakaumia:
  - diskreetti tasainen jakauma
  - binomijakauma
  - Poisson-jakauma
  - geometrinen jakauma

Keskeisiä jatkuvia jakaumia

- muuttujan arvot jatkuvia $\rightarrow$ jakauma jatkuva
- jatkuvan jakauman määrittää tiheysfunktio $f(X = x)$, jolle $\int_{-\infty}^{\infty} f(x)dx = 1$
- tiheysfunktion integraalifunktiota $F(x) = P(X \leq x)$ nimitetään jakaumafunktioksi (kertymäfunktioksi)
- keskeisiä jatkuvia jakaumia
  - tasainen jakauma, $X \sim Tas(a, b) \Leftrightarrow f(X = x) = \frac{1}{b-a}, a < x < b$
  - normaalijakauma
  - eksponenttijakauma
  - gammajakauma

Tasaisesti jakautuneet s-luvut

- *"If the numbers are not random, they are at least higgledy-piggledy ('sikin sokin')"*

  — GEORGE MARSAGLIA (1984)
- tietokoneen generoimat satunnaisluvut ovat pseudo-satunnaislukuja
- deterministisesti määräytyneitä, ei "oikeasti satunnaisia"
- mikä on riittävän satunnaista yhdelle sovellukselle ei välttämättä ole sitä toiselle

Tasaisesti jakautuneet s-luvut

- olemassa tilastollisia testejä, joilla testataan satunnaisuutta
- sama siemenluku $\rightarrow$ periodi alkaa samasta kohdasta $\rightarrow$ sama sarja "satunnaislukuja"
- käytännössä siemenluvun voi ottaa esim. systeemin kellonajasta
- testausvaiheessa on hyvä että siemenluvun voi asettaa käsin
- yleisin tasaisesti jakautuneiden satunnaislukujen generoimismenetelmä on *linear congruential method*

Linear congruential method

- m modulus
- a kertoja
- c lisäys
- X_0 alkuarvo
- satunnaislukusekvenssi saadaan rekursioyhtälöstä

$$X_{n+1} = (aX_n + c) \mod m, n \geq 0$$

- menetelmä on syklinen
- sykliä nimitetään periodiksi
- parametrien m, a, c valinta on erittäin keskeinen kysymys toimivuuden kannalta

Kertymäfunktio ja sen käänteisfunktio

- todennäköisyysjakauma voidaan siis ilmaista sen kertymäfunktion $F(x) = P(X \leq x)$ avulla
- $F(x)$ on aina monotoninen (kasvava) funktio
  - siis $F(x_1) \leq F(x_2)$, jos $x_1 \leq x_2$
  - F saa arvot välillä $[0, 1]$
- kun $F(x)$ on jatkuva ja aidosti kasvava, sillä on käänteisfunktio $F^{-1}(y), 0 < y < 1$
- $y = F(x)$ jos ja vain jos $x = F^{-1}(y)$
- käänteisfunktiota voidaan käyttää satunnaislukujen generoimiseen kyseisestä jakaumasta

Transformaatiomenetelmä

- olkoon $f(y)$ sen jakauman tiheysfunktio, josta halutaan generoida satunnaislukuja ja $F(y)$ vastaava kertymäfunktio
- oletetaan että kertymäfunktiolla on käänteisfunktio $y = F^{-1}(x), 0 < x < 1$
- transformaatiomenetelmä:
  1. generoi satunnaisluku $x \sim Tas(0, 1)$
  2. $y = F^{-1}(x)$ on satunnaisluku joka noudattaa haluttua jakaumaa

Esimerkki: eksponenttijakauma

- keskeisin jakauma tasaisen jakauman ja normaalijakauman jälkeen
- jos tulostimelle saapuu keskimäärin μ tulostuspyyntöä aikayksikköä kohti ja saapumisajat ovat satunnaisia, niiden välit noudattavat eksponenttijakaumaa

$$F(x) = 1 - e^{-x/\mu}, x > 0$$

- siis $F^{-1}(y) = -\mu \log(1 - y) \Rightarrow -\mu \log(1 - U) \sim \mathrm{Exp}(\mu)$
- koska $1 - U$ on tasaisesti jakautunut silloin kun U on

$\Rightarrow$ $x = -\mu \log U$ on eksponenttijakautunut odotusarvolla μ

Hylkäämismenetelmä

- *rejection sampling*
- myös hyväksymis-hylkäämismenetelmä
- ei edellytä kertymäfunktion eikä sen käänteisfunktion tuntemista
- on tunnettava apufunktio $g(x)$, joka saa kaikkialla suuremman arvo kuin $f(x)$ josta otoksia halutaan
- g:n alan oltava äärellinen

Algoritmi

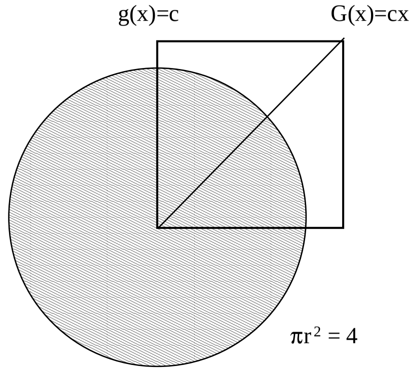
- olkoon $g(x) \geq f(x)$, kaikilla $x \in \Re$
- olkoon $A = \int_{-\infty}^{\infty} g(x)dx$ ja $G(x)$ vastaava integraalifunktio jolla on käänteisfunktio joka pystytään laskemaan
  1. generoi luku $y \sim Tas(0, A)$
  2. kandidaatti x saadaan transformaatiomenetelmellä $x = G^{-1}(y)$
  3. generoi $Z \sim Tas(0, g(x))$
  4. **if** $z < f(x)$ tulosta x **else goto** 1

Toinen muotoilu menetelmälle

- oletetaan että
  (i) h on sellaisen jakauman H tiheysfunktio, josta voidaan generoida arvoja
  (ii) $f(x) \leq c\, h(x)$, kaikille x, jollakin vakiolla c
- Algoritmi
  1. generoi kandidaatti $z \sim H$
  2. generoi $u \sim Tas(0, 1)$
  3. **if** $u \leq f(z)/(ch(z))$ **return** z; **else goto** 1
- tehokkuus riippuu c:stä

Hyvin simppeli esimerkki

- halutaan generoida satunnaislukuja jakaumasta, jonka tiheysfunktio on neljännesympyrän kehä



- $\pi r^2/4 = 1 \Rightarrow r = 2/\sqrt(\pi)$

Jatkoa

- olkoon apufunktio $g(x) = c, c > r \Rightarrow G(x) = cx$

1. generoidaan satunnaisluku $y \sim Tas(0, c^2)$
2. määritetään kandidaatti $x = G^{-1}(y) = y/c$
3. generoidaan satunnaisluku $u \sim Tas(0, g(x) = c)$
4. jos $f(x) = \sqrt{(\frac{4}{\pi} - x^2)} < u$ palautetaan x muuten palataan kohtaan 1

Tosielämässä

- monista keskeisistä jakaumista voidaan generoida satunnaislukuja hylkäämismenetelmää käyttäen
- esim. normaali-, gamma-, Poisson-, binomijakaumat
- keskeinen apufunktio g on Lorentzin jakauma tiheysfunktio; kyseisen jakauman integraalifunktio on tangenttifunktio

Arvojen riippumattomuus

- hylkäämismenetelmässä jokainen kandidaatti generoitiin siten, että peräkkäiset arvot eivät riippuneet toisistaan
- jos vakio c suuri, vain pieni osa “arvauksista” osuu oikeaan
- avaruus on liian “harva”, arvaukset osuvat liian harvoin
- hyvin monimutkaisista jakaumista (paljon ulottuvuuksia ja riippuvuuksia niiden välillä) ei riippumattomia arvoja voi generoida
- tilastollisessa mallintamisessa saatetaan tarvita otoksia hyvin “hankalista” jakaumista

MCMC-menetelmä

- Markovin ketju Monte Carlo
- sallitaan, että peräkkäiset arvot ovat toisistaan riippuvia
- kun on tehty hyvä “arvaus”, kokeillaan seuraavaksi jotakin lähellä olevaa arvoa
- joka kerta kun generoidaan arvoa halutusta jakaumasta, otetaan huomioon edellinen arvo, mutta ei muita

Markovin ketju

- *Markovin ketju* on jono satunnaismuuttujia siten, että jonon seuraava muuttuja riippuu edellisestä (mutta ei muista)

$$Pr(X_n | X_1, \ldots, X_{n-1}) \;=\; Pr(X_n | X_{n-1})$$

- idea: generoidaan ketju, jonka tilojen jakauma on otos halutusta jakaumasta

Algoritmi (Metropolis-Hastings)

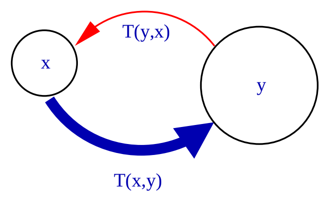
- valitse alkutila x_0
- toista K kertaa:
  - generoi kandidaatti x' jostakin jakaumasta $Q(x)$
  - hyväksy x' todennäköisyydellä $\alpha(x_n, x')$
  - muussa tapauksessa aseta $x_{n+1} := x_n$

Käänteisyysehto

- M-H -algoritmi perustuu *käänteisyysehtoon*
- jos kaikille pareille $x, y \in E$ pätee

$$f(x)\,T(x,y) \;=\; f(y)\,T(y,x)$$

$\Rightarrow$ ketjun tilojen jakauma on f



- kuinka valita siirtymätodennäköisyydet $T(x,y)$?

Milloin pitää hylätä?

- ilmaistaan käänteisyysehto seuraavasti:

$$f(x)\,q(x,x')\,\alpha(x,x') \;=\; f(x')\,q(x',x),$$

- $\alpha(x,x')$ on siis todennäköisyys, että kandidaatti hyväksytään
- jos $f(x)\,q(x,x') > f(x')\,q(x',x)$ hyväksytään aina
- muuten

$$\alpha(x,x') \;=\; \frac{f(x')\,q(x',x)}{f(x)\,q(x,x')}.$$

Hyväksymistodennäköisyys

- M-H -algoritmi edellyttää siis, että $\frac{f(x')}{f(x)}$ pystytään laskemaan
- käytännön kannalta sopivan ehdotusjakauman Q löytäminen tärkeää

Konvergenssi

- alkutila vaikuttaa tilojen jakaumaan
- jotta vaikutus eliminoitaisiin tarvitaan "lämmittelyjakso", jonka aikana ei vielä kerätä generoituja arvoja
- mikä on sopiva lämmittelyjakson pituus?
  - vaikea ongelma
- toinen keskeinen ongelma on, kuinka kauan simulaatiota pitää jatkaa, jotta otos olisi kattava

Yhteenveto

- satunnaislukujen generoiminen halutusta jakaumasta on tarpeen esimerkiksi Monte Carlo -testauksen yhteydessä
- tietokoneen generoimat satunnaisluvut ovat pseudo-satunnaislukuja
- tasaisesti jakautuneiden satunnaislukujen avulla voidaan generoida muita jakaumia noudattavia lukuja
  - transformaatiomenetelmä käyttää halutun jakauman kertymäfunktion käänteisfunktiota
  - hylkäämismenetelmässä käytetään apufunktiota josta osataan generoida otoksia ja joka "sulkee sisäänsä" halutun tiheysfunktion