Harjoitus 8: Monte-Carlo simulointi (Matlab)

SCI-C0200 Fysiikan ja matematiikan menetelmien studio 2024

Aalto University
School of Science

Systems
Analysis Laboratory

Harjoituksen aiheet

- Satunnaismuuttujien ja todennäköisyysjakaumien kertaus
- Tilastollinen estimointi
- Monte Carlo -simulointi

Oppimistavoitteet

- Ymmärtää periaatteessa kuinka Monte Carlo-simulointia voi käyttää systeemien analysoinnin tukena
- Monte Carlo -simuloinnin perusidea ja toteuttaminen Matlabilla

Satunnaismuuttujan todennäköisyysjakauma

- Satunnaismuuttuja X on satunnaisilmiön tulosvaihtoehtoihin liitetty reaaliarvoinen funktio $X : \Omega \to \mathbb{R}$. Se liittää reaaliluvun jokaiseen tapahtuma-avaruuden alkeistapahtumaan $\omega \in \Omega$.
  - Esim. Noppien silmälukujen summa, $X : \{(1,1), ..., (6,6)\} \to \{2, ..., 12\}$ s.e. $X(\omega) = \omega_1 + \omega_2$
- Todennäköisyysjakauma $P(X \in B)$ kertoo millä todennäköi- syydellä satunnaismuuttuja saa arvon joukosta B $\subset \mathbb{R}$.
  - $P(X = 2) = \frac{1}{36}$, $P(X \leq 3) = P(X = 2) + P(X = 3) = \frac{3}{36}$
- Satunnaismuuttujan X kertymäfunktio

$$F(x) = P(X \leq x)$$

kuvaa todennäköisyysmassan kertymistä argumentin x kasvaessa.

Tiheysfunktio $f(x)$

- Jatkuvat satunnaismuuttujat
  - Satunnaismuuttuja X on jatkuva, jos sen arvoalue on jokin reaaliakselin osaväli (tai jotkin osavälit).

$$f(x) = F'(x) \;\Rightarrow\; F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

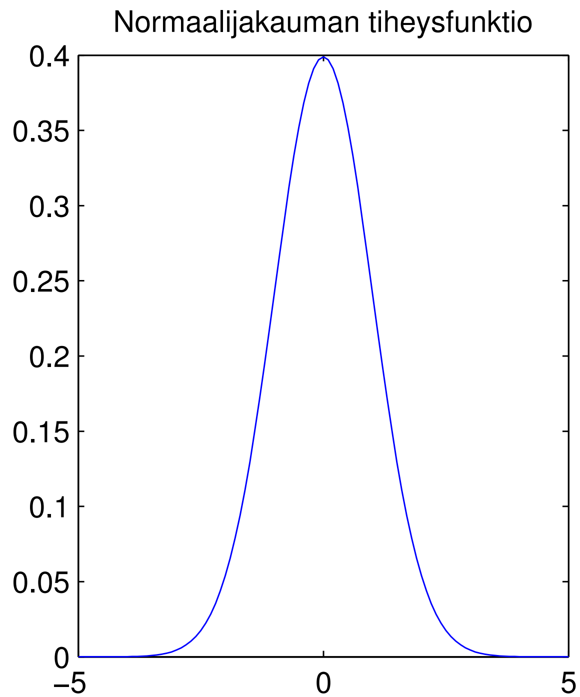
- Diskreetit satunnaismuuttujat
  - Satunnaismuuttuja X on diskreetti, jos sen arvoalue on diskreetti joukko eli muodostuu erillisistä reaaliakselin pisteistä

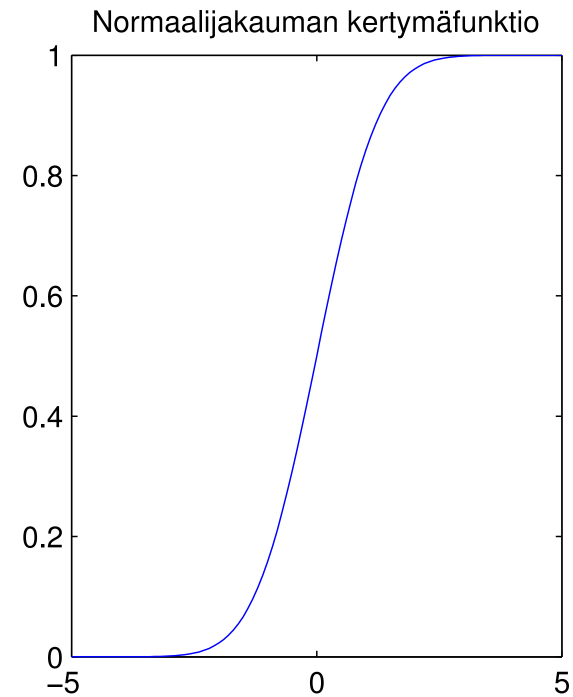
$$f(x) = P(X = x) \;\Rightarrow\; F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

  - $f(x)$:ää kutsutaan usein myös pistetodennäköisyysfunktioksi

Jatkuvan satunnaismuut. tiheys- ja kertymäfunktio

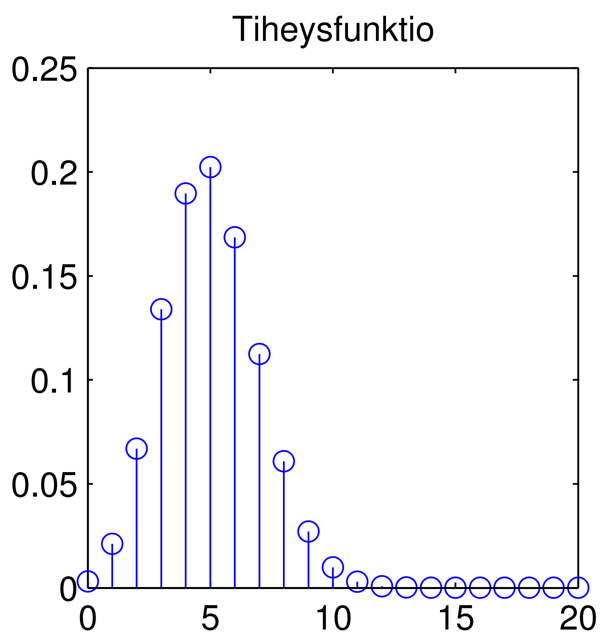
```
>> x = -5:0.1:5;
>> tiheys = normpdf(x,0,1); % N(0,1)-normaalijakauman tiheysfunktio
>> plot(x,tiheys)
>> kertyma = normcdf(x,0,1); % N(0,1)-normaalijakauman kertymäfunktio
>> plot(x,kertyma)
```

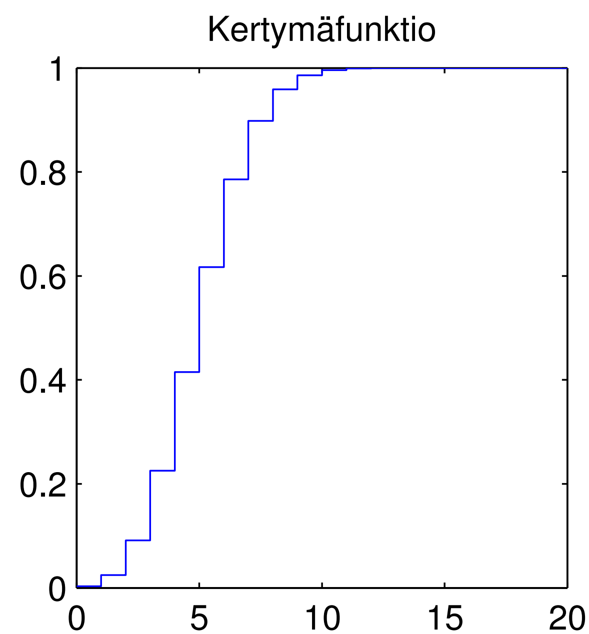




Diskreetin satunnaism. tiheys- ja kertymäfunktio

```
>> % Tehdään n=20 toistoa, onnistumistodennäköisyys p=0.25
>> k = 0:1:20;                     % Onnistumisia 0-20 kpl
>> tiheys=binopdf(k,20,0.25);   % binomijakauman tiheysfunktio
>> stem(k,tiheys);
>> kertyma=binocdf(k,20,0.25); % binomijakauman kertymafunktio
>> stairs(k,kertyma);
```





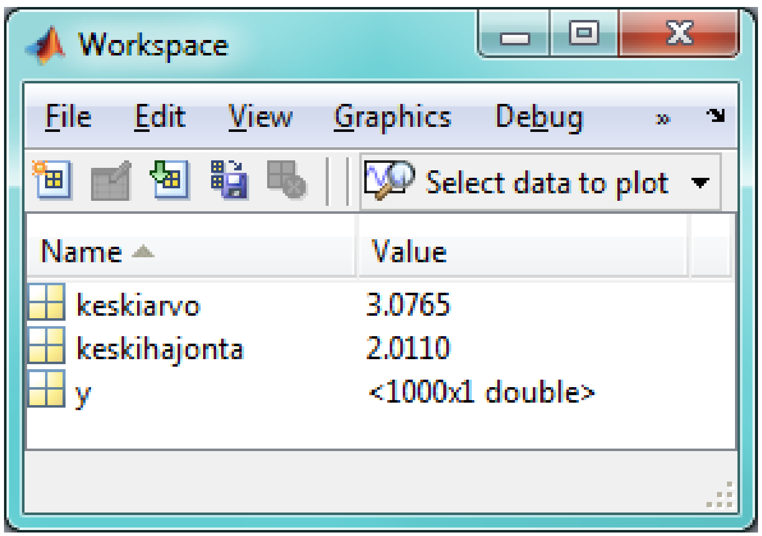
Tilastollinen estimointi

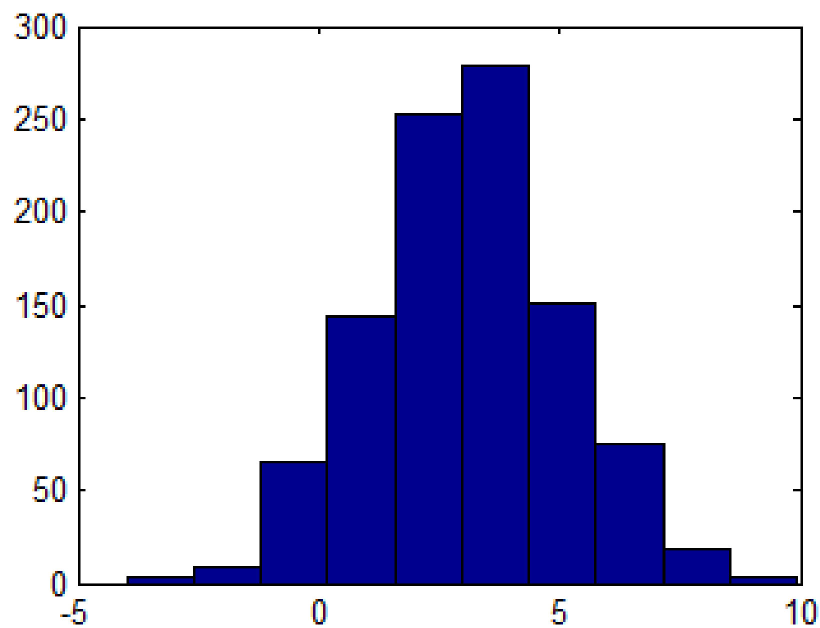
- Tilastollisen tutkimuksen tärkeimpiä osatehtäviä on estimoida eli arvioida havaintojen perusteella havainnot generoineen todennäköisyysjakauman tuntemattomat parametrit.
  - Esim. havaintojen keskiarvo on jakauman odotusarvon estimaatti (ks. Liitteet 1 ja 2)

- Havaittujen arvojen jakauma eli frekvenssijakauma voidaan esittää histogrammina (jatkuvat satunnaismuuttujat) tai pylväsdiagrammina (diskreetit satunnaismuuttujat)
  - Frekvenssi: esiintymiskertojen lukumäärä
  - Vastaa tiheys- tai pistetodennäköisyysfunktiota

Histogrammi jatkuvalle jakaumalle

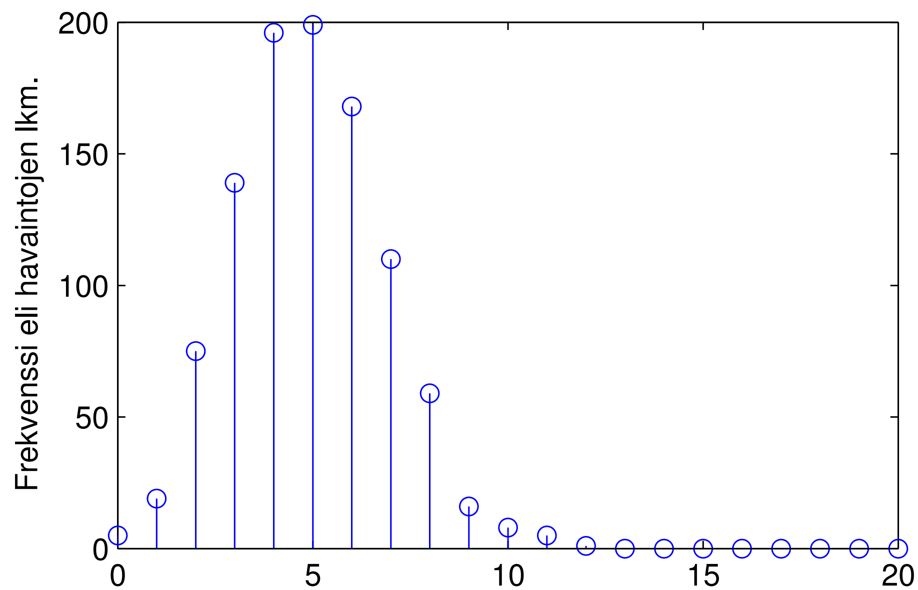
```
>> % Generoidaan 1000 satunnaislukua normaalijakaumasta, jonka
>> % odotusarvo on 3 ja varianssi 4
>> y=normrnd(3,2,1000,1);
>> % Muodostetaan histogrammi, lasketaan keskiarvo ja keskihajonta
>> hist(y); keskiarvo=mean(y); keskihajonta=std(y);
```





Pylväsdiagrammi diskreetille jakaumalle

```
>> % Generoidaan 1000 satunnaislukua binomijakaumasta, jossa
>> % toistojen lkm. n=20 ja onnistumistodennäköisyys p=0.25
>> y=binornd(20,0.25,1000,1);
>> % Muodostetaan pylväsdiagrammi ja lasketaan keskiarvo
>> stem([0:1:20],histc(y,[0:1:20])); mean(y)
ans = 4.9060
```



Monte Carlo -simulointi

Motivointi: Päätöksentekijä haluaa ymmärtää kuinka systeemin suorituskyky (performance / output) riippuu päätöksentekijän kontrollissa olevista muuttujista (decision variables / inputs). Lisäksi systeemiin liittyy satunnaisuutta.

Esimerkiksi: Kuinka monta kassaa tarvitaan pankkikonttoriin, jotta asiakkaat saadaan asiallisesti palveltua? Asiakasmäärä vaihtelee päivän mittaan satunnaisesti.

- Analyysiä varten tarvitaan matemaattinen malli tutkittavasta systeemistä sekä siihen liittyvistä satunnaisista muuttujista
- Monte Carlo simulointi hyvä työkalu kun syötteet (input) ja ulostulot (outputs) satunnaisia
- Ulostulon todennäköisyysjakauman päättely syötteen funktiona voi olla käytännössä liian vaikeaa ”analyyttisesti”

Monte Carlo -simulointi

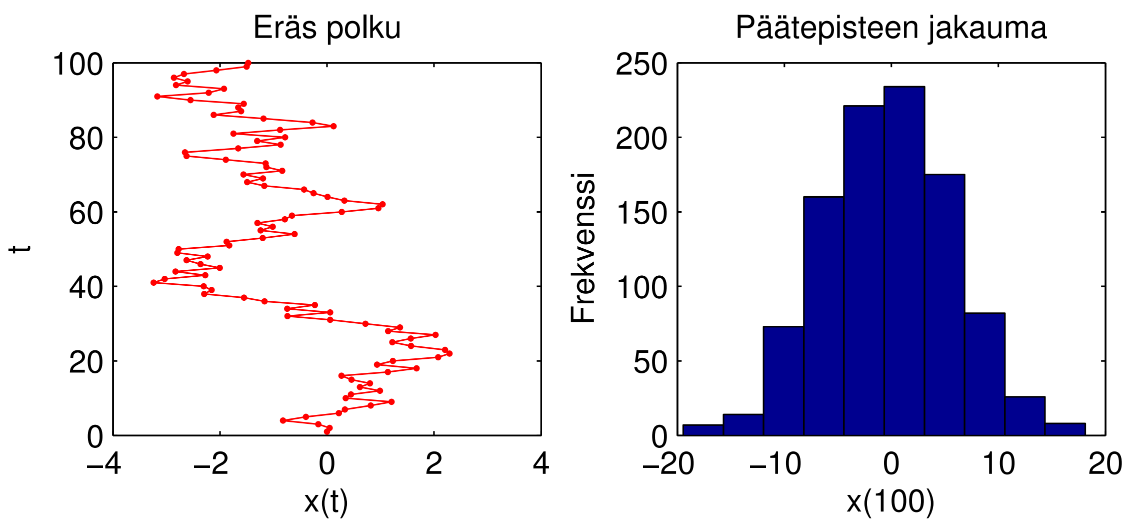
- Monte Carlo (MC) -simuloinnissa malli ajetaan useita (tuhansia) kertoja käyttäen syötteenä tietokoneen luomia satunnaislukuja

⇒ Saadaan tuhansia havaintoja ulostulosta. Näiden jakaumia voidaan havainnollistaa kuvin tai estimoida kiinnostavia tilastollisia tunnuslukuja.

- MC-menetelmiä käytetään hyvin laajasti esim. fysiikassa, tilastotieteissä, todennäköisyyslaskenntassa, laskennallisessa biologiassa, taloustieteissä, ...
- MC-menetelmiä varten tulee pystyä generoimaan satunnaislukuja halutusta todennäköisyysjakaumasta

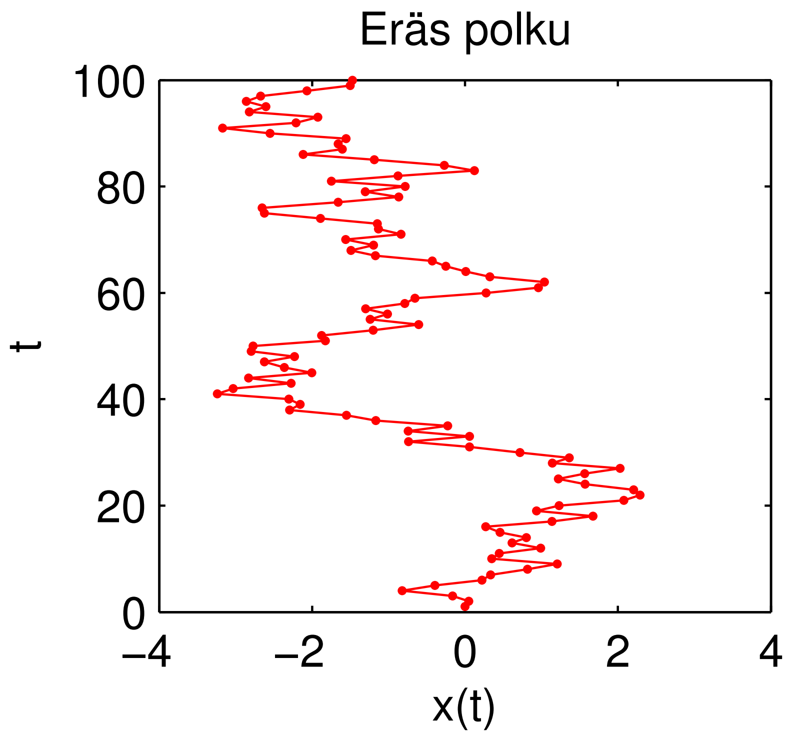
Monte Carlo –esimerkki: Satunnaiskävely

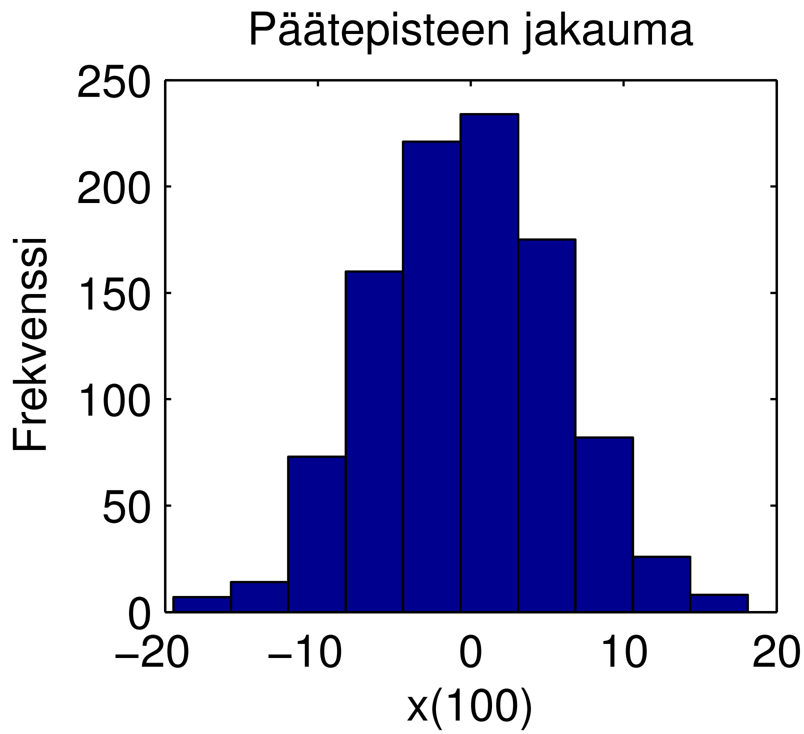
```
for mc=1:1000 % Tuhat kävelyä
  x(mc,1)=0; % Lähtöpiste origossa
  for t=2:100 % Simuloidaan yksi kävely
      x(mc,t)=x(mc,t-1)-1+2*rand; % x-siirtymä tasajakautunut [-1,1]
  end
end
subplot(1,2,1); plot(x(1,:),[1:1:100],'r.-'); title('Eräs polku')
subplot(1,2,2); hist(x(:,100)); title('Päätepisteen jakauma');
```



Monte Carlo –esimerkki: Satunnaiskävely

Satunnaiskävely on perusmalli esimerkiksi biologiassa (mikroskooppisten asioiden liike) ja investointiteoriassa (pörssikurssin kehitys)





Tehtävä A: Log-normaalijakauma

Moniin luonnonilmiöihin sekä teknisiin ja sosiaalisiin systeemeihin liittyy log-normaalisti jakautuneita suureita.
(`https://en.wikipedia.org/wiki/Log-normal_distribution`)

Jos satunnaismuuttuja X on normaalijakautunut, niin e^X on log-normaalijakautunut.

1. Generoi 500 lukua N(0,1)-normaalijakaumasta. Eli normaalijakauman parametreina odotusarvo 0 ja keskihajonta 1. Käytä `randn`-funktiota.
2. Tee luvuille eksponenttimuunnos. Esimerkki: Luvun 3 eksponenttimuunnos: `exp(3)=20.0855`.

🖩 Piirrä histogrammit sekä normaalijakautuneista luvuista sekä niiden eksponenttimuunnoksista. Käytä histogrammissa 20 luokkaa. ✍ Kuvaile histogrammeja sanallisesti.

3. Estimoi log-normaalijakauman parametrit käyttäen kohdassa 2 luotuja lukuja. Käytä funktiota `lognfit`. Selvitä parametreille 95%-luottamusvälit.

   (Parametrien tulisi olla lähellä arvoja 0 ja 1, sillä lähdimme liikkeelle N(0,1)-normaalijakaumasta.)

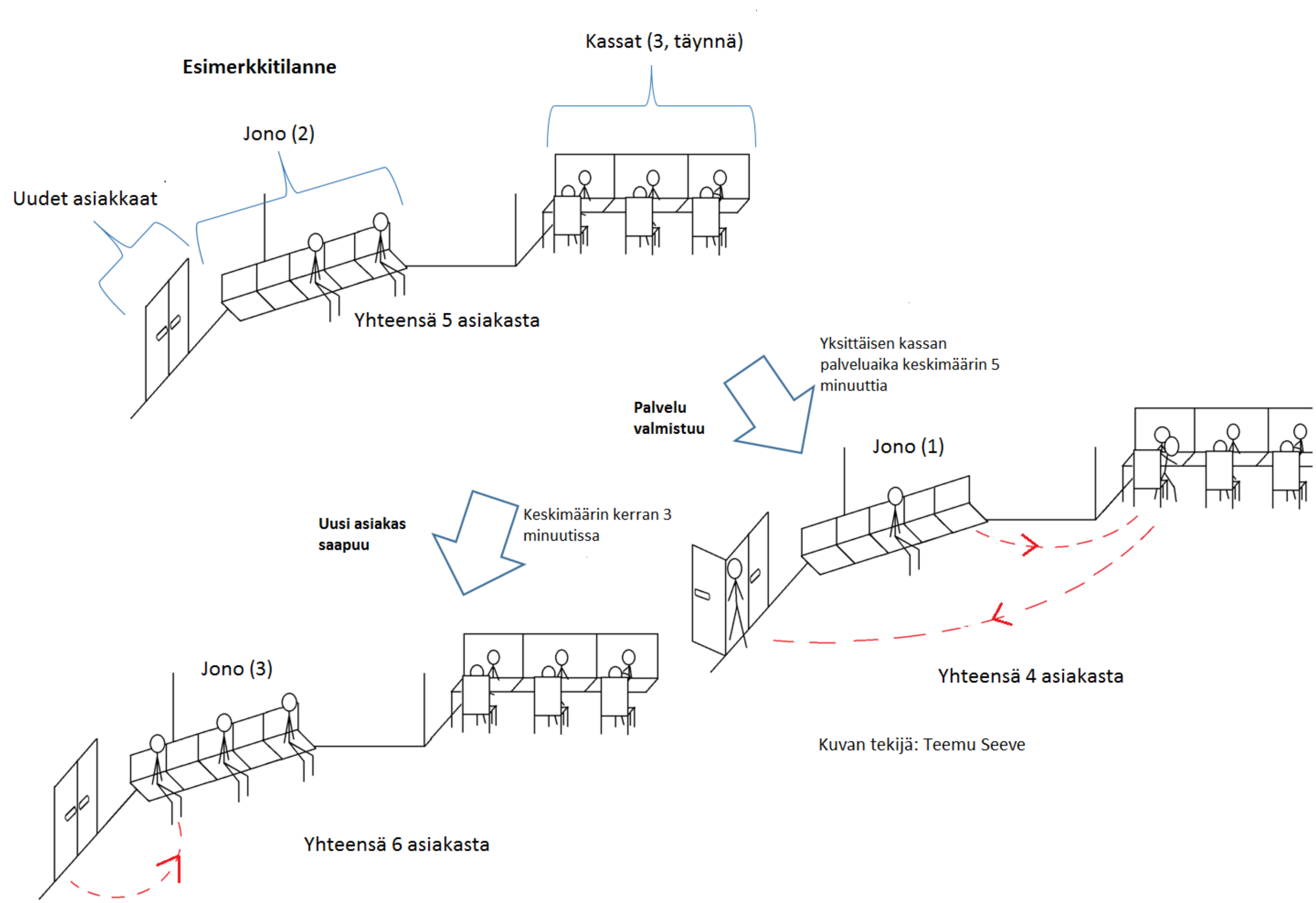
4. Toista työvaiheet 1, 2 ja 3 (kuvia ei tarvitse piirtää uudestaan). Generoi tällä kertaa vain 100 lukua.

✍ Miten parametreille estimoidut 95%-luottamusvälit muuttuvat, kun estimointiin käytetään pienempää otosta? Entäpä mitä käy parametreille estimoiduille 95%-luottamusväleille, kun estimointiin käytetään suurempaa otosta (generoi 10000 lukua)?

Tehtävä B: Palvelujonon Monte Carlo simulointi

Pankinjohtaja on palkannut sinut konsulttina selvittämään, kuinka monta kassaa pankissa tulisi olla, jotta jonon pituus pysyisi kohtuullisena.

- Asiakkaita saapuu pankkiin satunnaisesti, siten että kahden asiakkaan saapumisen välinen aika noudattaa eksponentiaalijakaumaa, jonka odotusarvo on 3 minuuttia (ks. Liite 3).
- Kullakin kassalla yhden asiakkaan palvelemiseen kuluva aika noudattaa exponentiaalijakaumaa, jonka odotusarvo on 5 minuuttia. Merkitään kassojen lukumäärää muuttujalla k.
- Saapuessaan pankkiin uusi asiakas siirtyy kassalle, jos sellainen on vapaana, tai jonoon, jos kaikkilla kassoilla on asiakas. Kun asiakkaan palvelu on valmistunut hän lähtee pois pankista.


Esimerkkitilanne
Kassat (3, täynnä)
Jono (2)
Uudet asiakkaat
Yhteensä 5 asiakasta
Yksittäisen kassan palveluaika keskimäärin 5 minuuttia
Palvelu valmistuu
Jono (1)
Uusi asiakas saapuu
Keskimäärin kerran 3 minuutissa
Yhteensä 4 asiakasta
Jono (3)
Kuvan tekijä: Teemu Seeve
Yhteensä 6 asiakasta

Tehtävä B: Lämmittelyä...

- Eksponentiaalisesti jakautuneita satunnaismuuttujia saat matlabista komennolla `exprnd`.

✍ Millä komennolla saat pystyvektorin, jossa on 15 eksponentiaalijakaumasta arvottua satunnaislukua? Käytä eksponentiaalijakauman odotusarvona lukua kolme.

1. While-rakenteen käyttö asiakkaiden kertymisen mallintamiseksi. Luo while-rakenne, jossa asiakkaat saapuvat yksitellen pankkiin ja kukaan ei poistu. Päivitä ajanhetkeä t kunnes 8 tuntia täyttyy.

```
while t<8*60 %Simulointi päättyy kun päivän kesto minuutteina ylittyy.
  t=t+...    %Päivitä uusi ajanhetki.
...
```

Piirrä kuva asiakkaiden lukumäärän kertymisestä.

Tehtävä B: Simuloinnin perusrakenne kun asiakkaita poistuu

Simuloinnin perusrakenne kun kassoja on useita ja asiakkaita poistuu:

- Pankin ovien avautuessa ajanhetkellä $t = 0$ minuuttia pankissa on $a = 0$ asiakasta.
- Ajanhetkeä t päivitetään käyttäen while-silmukkaa. Kassatilanteen simulointi jatkuu kunnes t ylittää päivän keston.
- **Jokaisella iteraatiolla arvotaan: (i) seuraavan uuden asiakkaan saapumisajanhetki sekä (ii) ajanhetket jolloin käytössä olevien kassojen palvelut valmistuvat.**
- Päivitä uudeksi ajanhetkeksi pienin näistä ajanhetkistä. Muut ajanhetket voi jättää huomiotta. Päivitä asiakasmäärärä.

Tehtävä B: Simulointi yhdellä kassalla

2. Simuloi ensin jonon pituutta kun kassoja on tasan yksi.

```
while t<8*60;
    %seuraavan asiakkaan saapuminen
    seuraava_asiakas_saapuu=

    %palveluiden valmistumiset
    kassoja_kaytossa=
    if kassoja_kaytossa > 0 %palvelu voi valmistua vain jos kassoja
                            %on käytössä
        palvelut_valmistuu=
    else %jos kassoja ei käytössä, niin palvelu ei valmistu koskaan
        palvelut_valmistuu=Inf;
    end
    %päivitä seuraavaksi ajanhetkeksi pienin näistä ajanhetkistä...
```

- Piirrä kuva jonon pituuden kehittymisestä. Käytä komentoa `stairs()`. Huom! kassoilla olevat asiakkaat eivät ole jonossa.
- Piirrä päällekkäni 10 aikasarjaa jonon kehittymisestä.

✍ Perustele parilla lauseella miksi *exp*-jakauman unohtamisominaisuus mahdollistaa ylläkuvatun simulointimallin käytön. (ks. Liite 3).

Tehtävä B: Simulointi usealla kassalla

3. Laajenna mallia toimimaan tapauksessa, jossa kassoja on useita. Tuo mukaan muuttuja k, joka kertoo kassojen lukumäärän.
4. Simuloi 10 jonon pituutta kuvaavaa (8h) aikasarjaa, kun kassoja on $k = 1, ..., 6$ kappaletta. Simulaation perusrakenne on esitetty kalvoilla 16 ja 17.

- Liitä vastauksiisi kuusi kuvaa (`subplot(3,2,k)`), kussakin 10 aikasarjaa tietyllä kassojen lukumäärällä. Nimeä akselit ja aseta otsikoksi kassojen lukumäärä (`strcat(num2str(k), ' kassaa')`).
- ✍ Laadi parin lauseen pituinen päätössuositus, jossa tuottamiisi kuviin tukeutuen perustelet optimaalisen kassojen lukumäärän pankinjohtajalle.
- ✍ Muuttaisitko päätössuositusta, jos toimeksiantajana olisi pankin pääluottamusmies?
- ✍ Lisää käyttämäsi Matlab-koodi vastauksiisi.

Kotitehtävä: Newsvendor-malli simuloiden

Tausta: Pankinjohtaja on kehunut osaamistasi kalakaverilleen Vesalle, joka tilaa sinulta uuden konsultointiprojektin.

1. Vesalla on vekottimien jälleenmyyntiyritys. Vekottimet tulevalle myyntijaksolle tilataan tehtaalta etukäteen tietämättä todellista kysyntää. Vekottimet ovat muotituotteita ja niillä ei ole käyttöä myyntijakson jälkeen: jos Vesa tilaa liikaa vekottimia, ylimääräiset heitetään roskiin. Jos Vesa tilaa liian vähän hän menettää potentiaalista myyntiä.

   - Kysyntä D on tasajakautunut välillä $[0, 300]$ kpl
   - Yhden vekottimen tilauskustannus on $c = 30$ euroa ja myyntihinta $p = 120$ euroa.

   ✍ Kirjoita arvauksesi optimaalisesta tilausmäärästä q.

Kotitehtävä: Newsvendor-malli simuloiden

2. Laadi Matlab-funktio, joka palauttaa `n` havaintoa myyntivoitosta annetulla tilausmäärällä q (VINKKI: generoi satunnainen kysyntä D ja käytä esim. `min`-komentoa laskemaan myyntivoitto)

   ✍ Mikä on todennäköisyys, että myyntivoitto on negatiivinen kun tilausmäärä on $q = 150$? (Vinkki: `find`)

3. Laadi Matlab-scripti, joka estimoi myyntivoiton odotusarvon tilausmäärillä $0, 1, ...300$ käyttäen 2.-kohdan funktiota. (Vinkki: Tässä tehtävässä riittävän tarkuuden saavuttamiseksi on syytä käyttää suurehkoa määrää simulaatioita. Esim. 10000 havaintoa kullakin q:n arvolla lienee riittävä määrä.)

   🖷 Liitä vastauksiisi kuva odotetusta myyntivoitosta tilausmäärän funktiona (nimeä akselit, otsikoksi oma nimi).

   ✍ Laadi parin lauseen pituinen päätössuositus, jossa kuvaan tukeutuen perustelet optimaalisen tilausmäärän Vesalle.

Kotitehtävä: Newsvendor-malli

4. Vesa on huomannut että tehdas ei aina toimita kaikkia hänen tilaamiaan vekottimia. Laajenna malliasi siten, että se estimoi myyntivoiton odotusarvon eri tilausmäärillä kun toimitettujen vekottimien osuus tilausmäärästä Z on tasajakautunut välille $[0, 1]$ (Vinkki: Voit käyttää esim. `rand` -komentoa). Vesa tietysti maksaa vain toimitetuista vekottimista.

- Maksimin löytämiseksi täytyy tässä kohdassa laajentaa tilausmäärä -akselia. Estimoi myyntivoiton odotusarvo vaikkapa tilausmäärillä $0, 1, ..., 500$.

- Liitä vastauksiisi kuva odotetusta myyntivoitosta tilaus määrän funktiona (nimeä akselit, otsikoksi oma nimi).
- ✍ Laadi parin lauseen pituinen päätössuositus, jossa kuvaan tukeutuen perustelet optimaalisen tilausmäärän Vesalle.
- ✍ Vertaa simuloimalla saatua suositusta 6. harjoituksissa laskemaasi tulokseen.

Liitä vastauksiisi kaikki lähdekoodisi kommentoituna.

Liite 1: Odotusarvon estimointi

- Satunnaismuuttujan X odotusarvo on sen jakauman painopiste
- Diskreetille jakaumalle odotusarvo määritellään seuraavasti:

$$\mathrm{E}(X) = \mu_X = \sum_i x_i p_i = \sum_i x_i f(x_i)$$

- Vastaavasti jatkuvalle jakaumalle odotusarvo määritellään näin:

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{+\infty} x f(x) dx$$

- Odotusarvon estimaatti on havaintoaineiston aritmeettinen keskiarvo:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Liite 2: Varianssin ja keskihajonnan estimointi

- Satunnaismuuttujan X varianssi on satunnaismuuttujan odotusarvosta lasketun poikkeaman neliön odotusarvo.
- Diskreetin satunnaismuuttujan varianssi:

$$\mathrm{Var}(X) = \sigma_X^2 = \sum_i (x_i - \mu_x)^2 p_i = \sum_i (x_i - \mu_x)^2 f(x_i)$$

- Jatkuvan satunnaismuuttujan varianssi:

$$\mathrm{Var}(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x) dx$$

- Standardipoikkeama eli keskihajonta $\sigma_X = \sqrt{\mathrm{Var}(X)}$
- Varianssin estimaatti on havaintoaiheinton otosvarianssi:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

Liite 3: Eksponenttijakauma

- T on eksponentiaalijakautunut intensiteetillä λ (eli $T \sim \exp(\lambda)$)
  - Kertymäfunktio: $F(t) = P(T \leq t) = 1 - e^{-\lambda t}$ kaikilla $t \geq 0$.
  - Tiheysfunktio: $f(t) = \frac{\partial}{\partial t} P(T \leq t) = \lambda e^{-\lambda t}$
  - $\mathbb{E}[V] = \frac{1}{\lambda}$, $\mathrm{Var}[V] = \frac{1}{\lambda^2}$
  - $P(T > t) = 1 - F(t) = e^{-\lambda t}$
- Unohtamisominaisuus: Olkoon $T \sim \exp(\lambda)$ ajanhetki, jolloin pankkin saapuu seuraava asiakas. Kun on kulunut s minuuttia ilman että uusia asiakkaita on saapunut, aika seuraavan asiakkaan saapumiseen on edelleen $\exp(\lambda)$ jakautunut, sillä

$$P(T > s + t | T > s) = \frac{P(T > s + t, T > s)}{P(T > s)} = \frac{P(T > s + t)}{P(T > s)}$$

$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = \frac{e^{-\lambda s} e^{-\lambda t}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t)$$

Liite 3: Eksponenttijakauma

- Unohtamisominaisuudesta enemmän kiinnostuneet voivat syventyä edellisten laskujen perusteluihin tulkitsemalla Bayesin teoreemaa:

$$P(T > s + t \mid T > s) = \frac{P(T > s \mid T > s + t)P(T > s + t)}{P(T > s)}$$

  Kun arvot s ja t ovat positiivisia, niin on selvää, että $s + t > s$. Siksi todennäköisyys tapahtumalle $T > s$ ehdolla $T > s + t$ on 1. Toisin sanoen, $P(T > s|T > s + t) = 1$.

- Edellisen voi tulkita sanallisesti seuraavasti: on varmaa, että satunnaismuuttuja saa aina suurempia arvoja kuin arvo s, jos satunnaismuuttujalla on ehtona saada suurempia arvoja kuin $s + t$.
- Tämä huomio selittää ensimmäiset kaksi yhtäsuuruutta edellisen sivun laskuissa.