Johdatus todennäköisyyslaskentaan
Kevät 2014

Luento 13 / 13

Jukka Kohonen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

Summan tai keskiarvon jakauma

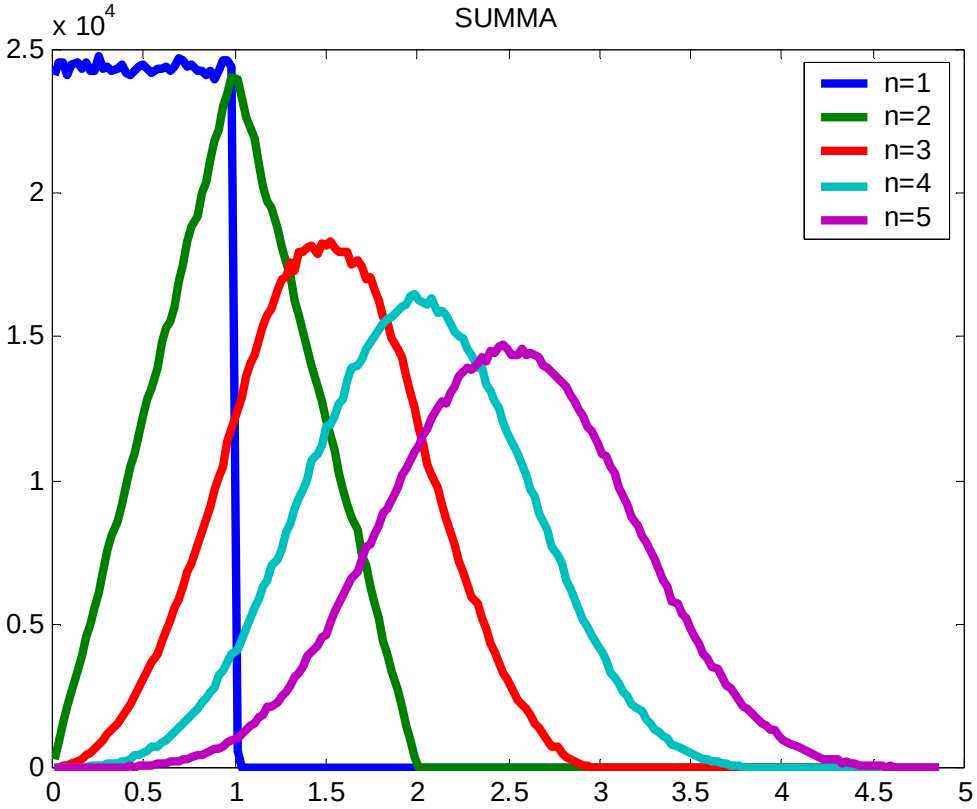
Olkoot $\boldsymbol{X_1,\ldots,X_n}$ riippumattomia, samoin jakautuneita, jostakin jakaumasta.
(Esim. Tas, Exp, Geom, Bernoulli, nopanheitto, ...)
Olkoon $E(X_i)=\boldsymbol{\mu}$ ja $D(X_i)=\boldsymbol{\sigma}$.

Mitä tiedetään summasta $\boldsymbol{S}=X_1+\ldots+X_n$ ja/tai keskiarvosta $\boldsymbol{M} = S/n$?

- Helppoa: $\mathbf{E}(\boldsymbol{M}) = \boldsymbol{\mu}$ (odotusarvojen yhteenlasku)
- Melkein helppoa: $\mathbf{D}(\boldsymbol{M}) = \boldsymbol{\sigma} / \sqrt{\boldsymbol{n}}$ (varianssien yhteenlasku)
- Vähän vaikeampi: $\boldsymbol{M \approx \mu}$ **todennäköisesti** (suurten lukujen laki)
- Vaikeampi: **Mikä on *M*:n jakauman muoto**?
  ➔ keskeinen raja-arvolause

KESKEINEN RAJA-ARVOLAUSE

Tas(0,1)-muuttujien summa



Yksittäisen muuttujan
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

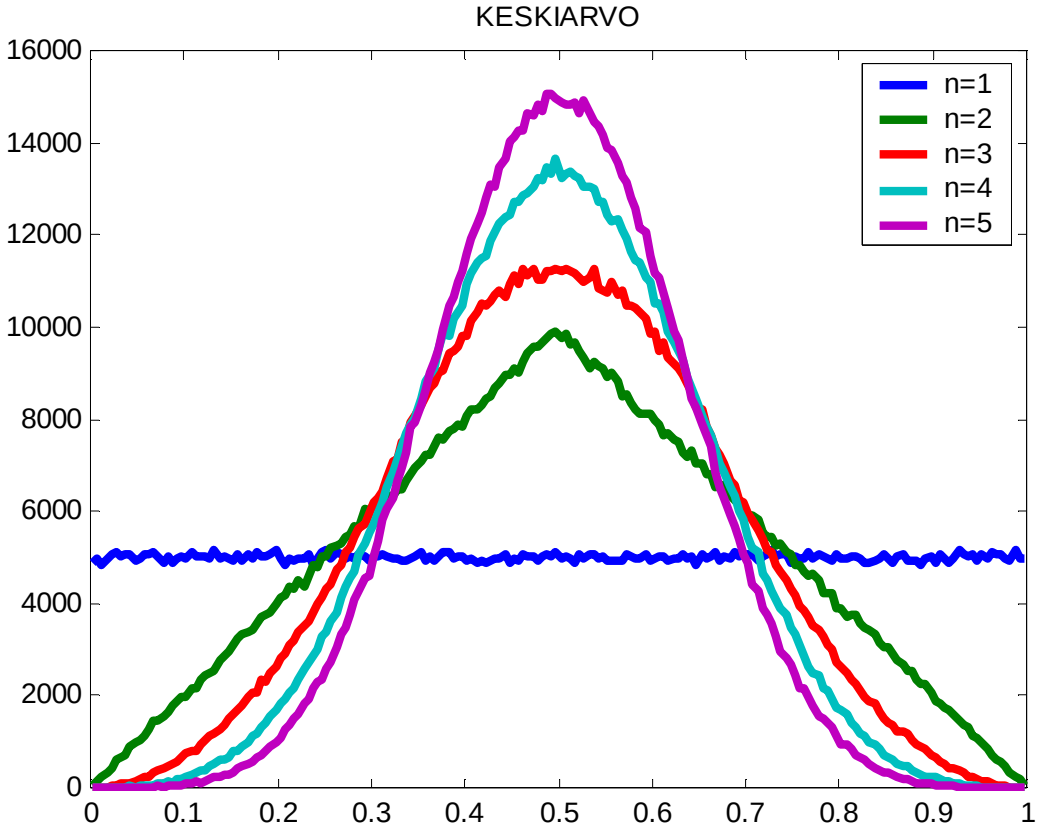
Summa
$\mathbf{S_n = \Sigma X_i}$

Summan tiheysfunktio on hankala kokoelma (*n*-1):nnen asteen polynomeja, jotka voi laskea konvoluutiolla. Kuvassa empiirinen histogrammi.

Kun n kasvaa, **summan** jakauma

- siirtyy oikealle: $E(S_n) = \mu \cdot \mathbf{n}$
- **levenee**: $D(S_n) = \sigma \cdot \sqrt{\mathbf{n}}$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Tas(0,1)-muuttujien keskiarvo



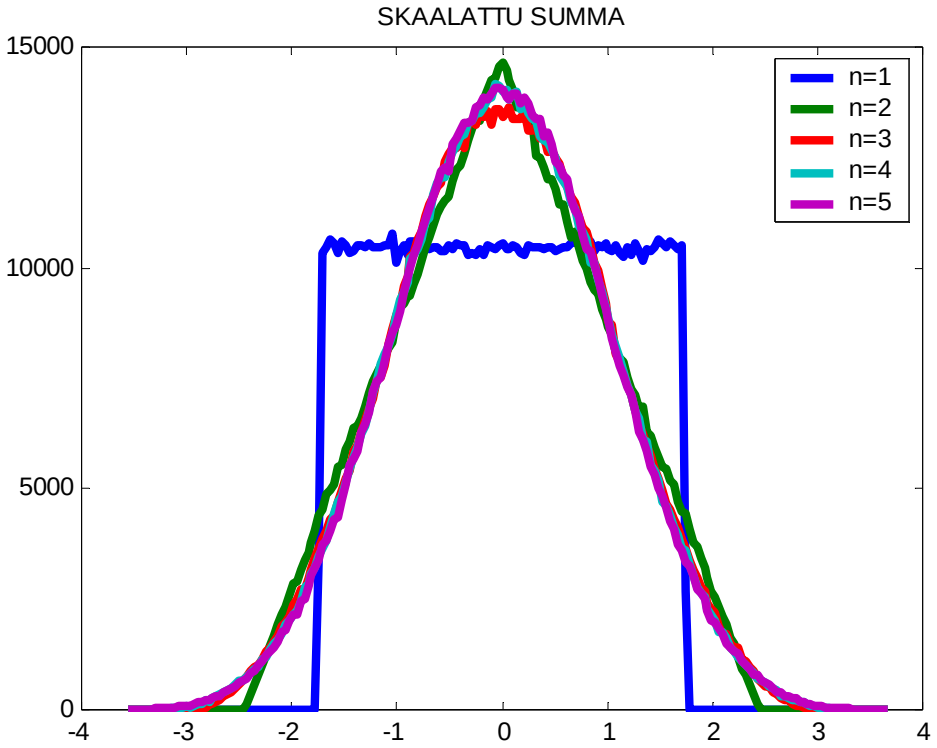
Yksittäisen muuttujan
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

(Otos)keskiarvo
$M_n = S_n / n$

Kun n kasvaa, **keskiarvon** jakauma

- pysyy paikallaan: $E(M_n) = \mu = 0.5$
- **kapenee**: $D(M_n) = \sigma / \sqrt{n}$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Tas(0,1)-muuttujien skaalattu summa



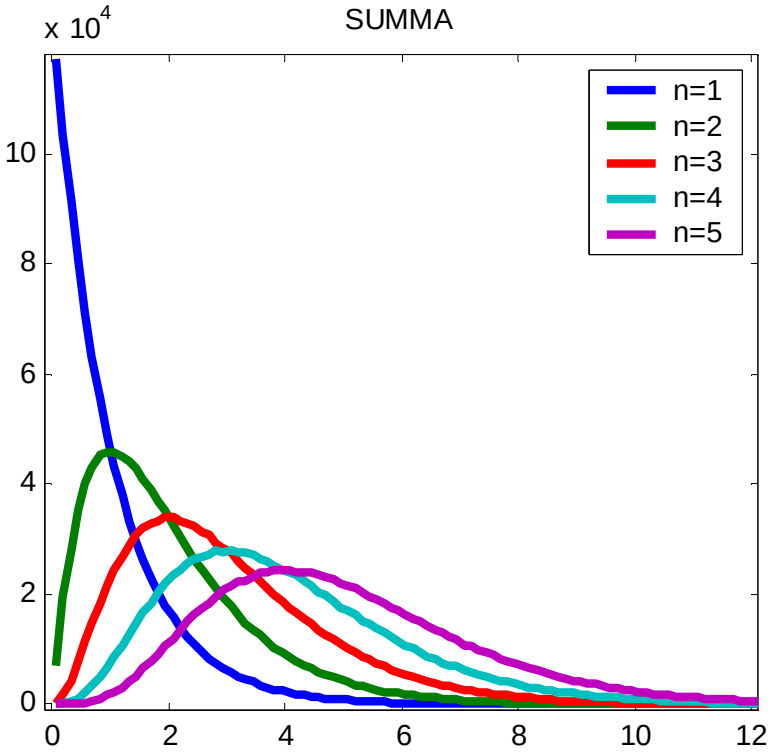
Yksittäisen muuttujan
$\mu = E(X_n) = ½$
$\sigma = D(X_n) = 1 / \sqrt{12}$

Skaalattu summa
$\mathbf{Z_n = [S_n - E(S_n)] / D(S_n)}$

Kun n kasvaa, **skaalatun summan** jakauma

- pysyy paikallaan: $E(Z_n) = 0$
- pysyy saman levyisenä: $D(Z_n) = 1$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Exp(1)-muuttujien summa



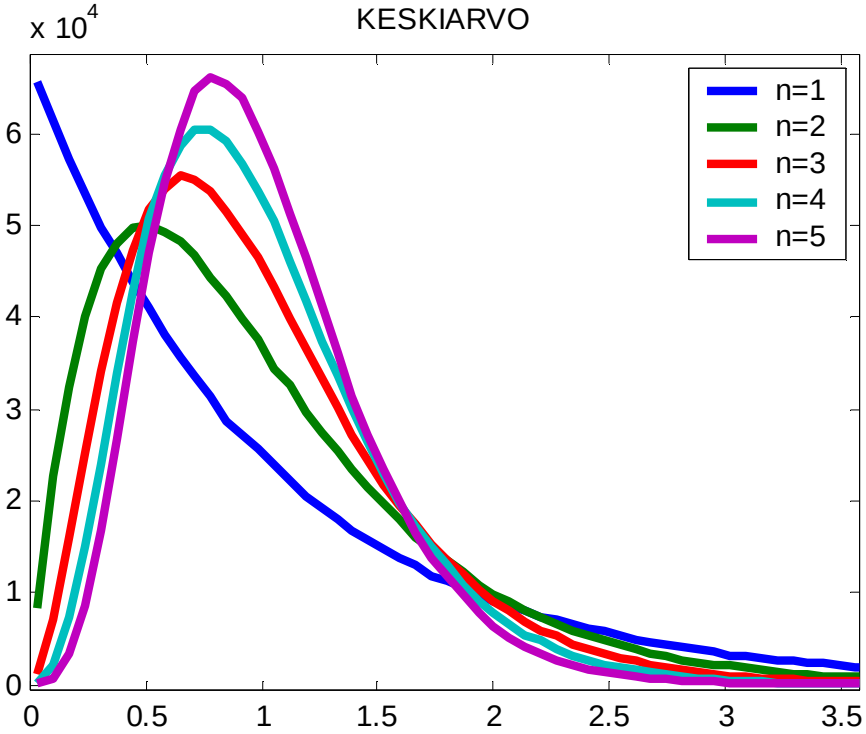
Yksittäisen muuttujan
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

Summa
$\mathbf{S_n = \Sigma X_i}$

Summan jakauma on ns. gammajakauma (Tuominen s.106-109)

Kun n kasvaa, **summan** jakauma
- siirtyy oikealle: $E(S_n) = \mu \cdot \mathbf{n}$
- **levenee**: $D(S_n) = \sigma \cdot \sqrt{\mathbf{n}}$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Exp(1)-muuttujien keskiarvo



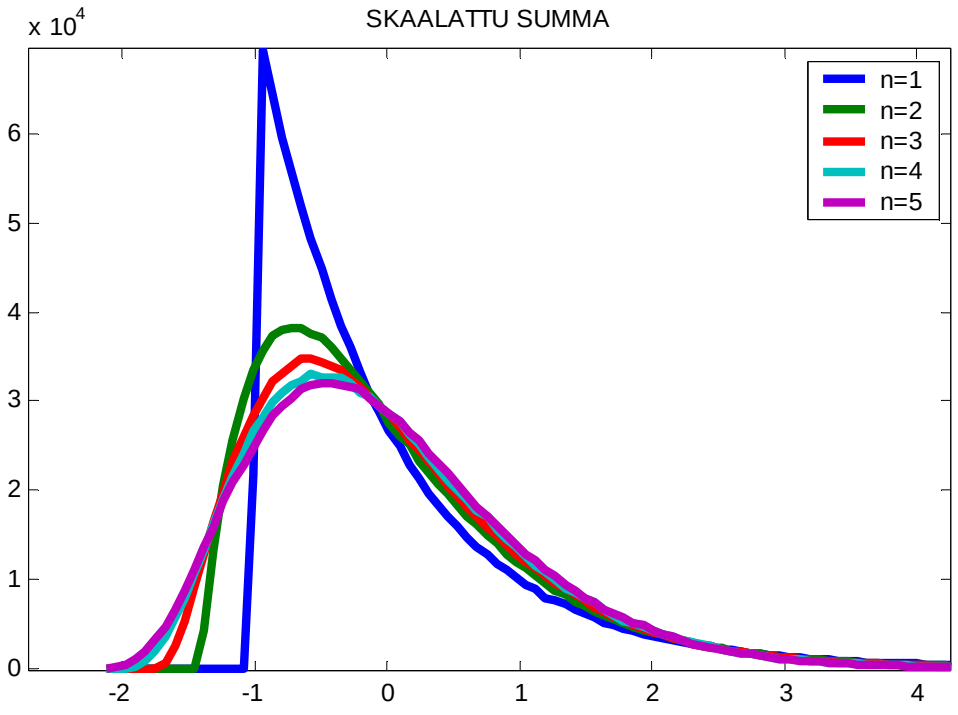
Yksittäisen muuttujan
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

(Otos)keskiarvo
$\mathbf{M_n = S_n / n}$

Kun n kasvaa, **keskiarvon** jakauma

- pysyy paikallaan: $E(M_n) = \mu = 1$
- **kapenee**: $D(M_n) = \sigma / \sqrt{n}$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Exp(1)-muuttujien skaalattu summa



Yksittäisen muuttujan
$\mu = E(X_n) = 1$
$\sigma = D(X_n) = 1$

Skaalattu summa
$\mathbf{Z_n = [S_n - E(S_n)] / D(S_n)}$

Exp on hyvin epäsymmetrinen jakauma. Vielä 5 muuttujan summassa näkyy melkoinen epäsymmetria, ei kovin lähellä normaalijakaumaa.

Kun n kasvaa, **skaalatun summan** jakauma
- pysyy paikallaan: $E(Z_n) = 0$
- pysyy saman levyisenä: $D(Z_n) = 1$
- muuttuu muodoltaan lähemmäs normaalijakaumaa

Empiirinen havainto

Kun **riippumattomia** satunnaismuuttujia lasketaan yhteen, **summan jakauma näyttää muodoltaan** useimmiten melkein samalta.

Kyseessä on jakaumaperhe nimeltään ”**normaalijakauma**”, merk. ***N***

(**Perhe** tarkoittaa, että on olemassa monta eri normaalijakaumaa, aivan kuten on monta eri Tas-jakaumaa ja monta eri Exp-jakaumaa; **parametrit** osoittavat perheestä yhden tietyn jakauman.)

Tällä kurssilla:

- Määrittelemme normaalijakauman
- Toteamme sen ominaisuuksia
- Tyydymme em. empiiriseen havaintoon
- Käytämme sitä hyväksi, kun **approksimoimme summan jakaumaa**

Havainto on formaalisti nimeltään *keskeinen raja-arvolause*, ja se pystytään kyllä todistamaankin (Tuominen s. 118).

Normaalijakauma Tuominen 61-64

- Standardinormaalijakauma on eräs jatkuva jakauma
  - tiheysfunktio: $f(x) = c \cdot \exp(-0.5x^2)$ Emme välitä *c*:stä nyt

- Lausekkeesta nähdään mm.
  - Tiheys suurimmillaan kohdassa $x=0$ (miksi?)
  - Jakauma on symmetrinen kohdan $x=0$ suhteen
  - Tiheys pienenee *hyvin nopeasti*, kun |x| kasvaa (neliön eksponenttifunktio!)
  - Mediaani ja odotusarvo ovat = 0
    (tarvitaan vähän päättelyä)

Normaalijakauma

- Kuten tasajakaumalle jne., myös standardinormaalijakaumalle voidaan tehdä muunnoksia. Saadaan uusi jakauma.
  - tärkeitä ovat vakiolla kertominen ja vakion lisääminen: näissä muoto pysyy samana, mutta jakauman paikka tai leveys muuttuu.
- Määrittelemme Tuomisen (s. 62) tapaan:

  Jos Z on standardinormaalijakaunut ja

  $$\boldsymbol{X = aZ + b},$$

  niin X on normaalijakautunut tällaisin parametrein:

  $$\boldsymbol{X \sim N(b, a^2)}$$

Normaalijakauma

- Kertymäfunktiota tarvitaan, kun lasketaan välien todennäköisyyksiä (Tuominen s. 63)
- Kertymäfunktio pitäisi saada integroimalla tiheysfunktiota. Ikävä kyllä integraalille ei saada nättiä lauseketta (Huomautus 2.3.10 s. 61)
- Ratkaisu:
  - käytetään kertymäfunktion taulukkoa tai
  - käytetään laskukonetta `normcdf`

Normaalijakauma

- Olkoon Z standardinormaalijakautunut.
  - Edellä päättelimme $E(Z) = \mathbf{0}$
  - Voidaan laskea $D(Z) = \mathbf{1}$ (Tuominen s. 85)
- Jos nyt

  $X = aZ + b$,

  osaamme päätellä

  $E(X) = a \cdot 0 + b = b$

  $D(X) = a \cdot 1 = a$

- Normaalijakauma saadaan siis haluttuun kohtaan (odotusarvo b) ja halutun levyiseksi (hajonta a).
- Jakaumassa $N(b, a^2)$ ensimmäinen parametri siis kertoo odotusarvon ja jälkimmäinen varianssin. Tavallisesti niille käytetään symboleja μ ja σ^2

Normaalijakaumien summa

Oletetaan, että $X \perp\!\!\!\perp Y$, ja

$X \sim N(\mu_1, \sigma_1^2)$

$Y \sim N(\mu_2, \sigma_2^2)$

Tällöin myös **summa on normaalijakautunut** (Tuominen s. 73)

X+Y ~ N(jollain parametreilla). Mutta millä parametreilla?

Muistetaan odotusarvon ja varianssin summakaavat:

E(X+Y) = E(X) + E(Y)

Var(X+Y) = Var(X) + Var(Y) (riippumattomuus)

Parametrit on siis helppo päätellä

$X+Y \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$

Odotusarvot summataan ja varianssit summataan (kun $X \perp\!\!\!\perp Y$).

Normaalijakaumien summa

Kahden eri bussilinjan ajoajat (riippumattomasti)

Bussi 1: $X \sim N(20, 4^2)$

Bussi 2: $Y \sim N(24, 4^2)$

Matkat tehdään peräkkäin. Koko matka-aika

$$X+Y \sim N(20+24, 4^2+4^2)$$
$$= N(44, 5.66^2)$$

Huom: Hajonta ei kasvanut kovin paljon (ei siis ”4 + 4 = 8 min”). *Varianssit* laskettiin yhteen ja hajonta on varianssin neliöjuuri.

Mikä on tn, että matka-aika < 50 min?

$$F_{X+Y}(50) = \Phi((50-44) / 5.66) = 0.855$$

Normaalijakaumien erotus

Kahden eri bussilinjan ajoajat (riippumattomasti)

Bussi 1: $X \sim N(20, 4^2)$

Bussi 2: $Y \sim N(24, 4^2)$

Bussit lähtevät samaan aikaan eri paikoista, ja saapuvat em. ajassa samalle pysäkille. Herra K on ykkösbussin kyydissä ja haluaa vaihtaa kakkosbussiin.

Mikä on vaihtamiseen jäävän ajan $V = (Y - X)$ jakauma?

Huomataan, että V on normaalijakautuneiden summa: $V = Y + (-X)$,

missä $-X \sim N(-20, 4^2)$ (normaalijak. kertominen vakiolla −1)

Siis $V \sim N(24-20, 4^2+4^2)$

$= N(4, 5.66^2)$

Odotusarvot vähennettiin toisistaan (ei yllätys) mutta varianssit summattiin.

Vaihtoaikaa on keskimäärin 4 min. Hajonta 5.66 on siihen nähden melko suuri.

Mikä on tn, että vaihtoaika on negatiivinen (jolloin vaihto epäonnistuu)?

$$P(V < 0) = F_V(0) = \Phi((0-4) / 5.66) = 0.24$$

Keskeinen raja-arvolause (1/2)

- Olipa yksittäisten muuttujien X_i jakauma mikä tahansa (kunhan odotusarvo ja varianssi on olemassa, ja muuttujat riippumattomia), niin summan (ja keskiarvon) jakauma ”normaalistuu”.

- Koska otossumma ja otoskeskiarvo ovat ”liikkuvia maaleja”, formaali raja-arvotulos esitetään *skaalatulle summalle,* jonka odotusarvo=0 ja hajonta=1 pysyvät paikallaan *n*:n kasvaessa

- Ilmiö, muodon normaalistuminen, tapahtuu kuitenkin **aivan samalla tavalla myös otossummalle ja otoskeskiarvolle**.

Keskeinen raja-arvolause (2/2)

Formaali raja-arvotulos esitetään standardoidun summan **kertymäfunktiolle:** jokaisessa pisteessä $b \in \mathbb{R}$ pätee

$$P(Z_n \leq b) \rightarrow \Phi(b), \quad \text{kun } n \rightarrow \infty$$

- Pätee yhtä hyvin jatkuville ja diskreeteille sm:ille

- Kertymäfunktiosta saadaan suoraan se, mitä useimmiten halutaan eli välin tn:

  $$P(a < Z_n \leq b) = F(b) - F(a),$$

  jossa kertymäfunktion arvot KRL:n mukaan lähestyvät standardinormaalin kertymäfunktion arvoja Φ(b) ja Φ(a).

  Siksi arvioidaan

  $$P(a < Z_n \leq b) \approx \Phi(b) - \Phi(a).$$

KRL:n käyttäminen

- Käytännössä ei tarvitse käyttää standardoitua summaa: voidaan käyttää suoraan sm:ien **summan tai keskiarvon jakaumaa, joka approksimoidaan normaaliksi.**

- Tarvitaan tietysti jakauman **parametrit**, mutta ne on **helppo päätellä** odotusarvon ja varianssin summakaavoilla. Esim. $E(S_n) = n\,E(X_i)$, koska summataan n termiä.

- Kun arvioidaan, että summassa $S_n = X_1+\ldots+X_n$ on **tarpeeksi termejä** (*ja KRL:n muut ehdot täyttyvät*), approksimoidaan että summa (keskiarvo) on normaalijakautunut ja lasketaan halutut todennäköisyydet.

Mikä on ”tarpeeksi termejä”?

Riippuu

- Alkuperäisen jakauman **muodosta**: esim. hyvin vino jakauma (Exp) normaalistuu hitaammin kuin symmetrinen (Tas tai kolikonheitto tai noppa).
- **Mitä kohtaa** jakaumasta approksimoidaan. Normaaliapproksimaatio on tarkempi jakauman keskellä ja huonompi hännissä.
- Erityisesti: summalla on useinkin ehdoton **alaraja ja/tai yläraja** (esim. Tas ja Exp), jonka ulkopuolella todellinen tn on nolla ja normaaliapproksimaatio pielessä.
- Tarkkaa nyrkkisääntöä vaikea asettaa, mutta noin 20 muuttujan summalla normaalijakauma on ”yleensä” hyvin tarkka (paitsi aivan jakauman hännissä).

Paristoesimerkki

- Käytetään peräjälkeen *n*=20 paristoa, käyttöajan odotusarvo μ=½ vuotta ja hajonta σ=½ vuotta, riippumattomia. Jakauman muotoa emme tunne
- Koko käyttöikä S_{20} = käyttöikien summa.

- Summakaavojen mukaisesti

  $E(S_{20}) = 20 \cdot \mu = 10$ vuotta

  $D(S_{20}) = (\sqrt{20}) \cdot \sigma = 2.236$ vuotta

- Oletetaan S_{20} normaalijakautuneeksi näillä parametreilla:

  $S_{20} \sim N(10, 2.236^2)$

- Tn, että kestää alle 15 vuotta:

  $F(15) = \Phi((15-10) / 2.236) \approx 0.987$

Kolikkoesimerkki

- Heitetään miljoona kolikkoa.
- Kruunien lkm $S \sim \mathrm{Bin}(10^6, ½)$
- Tiedetään $E(S) = 500000$
  $D(S) = 500$
- Likimain $S \sim N(500000, 500^2)$
- Nyt tn. että lukumäärä poikkeaa odotusarvosta enintään tuhannella (eli kahdella hajonnalla)

$$P(-1000 \le S - E(S) \le 1000) \approx \Phi(2) - \Phi(-2) \approx 0.955$$

- Tarkankin arvon voisi laskea summaamalla 2001 kpl binomijakauman pistetodennäköisyyksiä (joissa on aika isoja binomikertoimia).

Kolikkoesimerkki: Normaali vs. Tsebysev

- Normaalijakaumalla saimme

  $P(-1000 \le S - E(S) \le 1000) \approx \Phi(2) - \Phi(-2)$ **≈ 0.955**

- Jos emme uskaltaisi olettaa summaa normaalijakautuneeksi, niin pelkän odotusarvon ja hajonnan perusteella voisimme soveltaa Tsebyseviä (häntätodennäköisyys 2 hajonnalle eli k=2):

  $P(|S - E(S)| \ge 1000) \le 1/k^2 = 0.25$

  $P(|S - E(S)| \ge 1000)$ **≥ 0.75**

  Raja on paljon varovaisempi, mutta se ainakin **varmasti pitää paikkansa** eikä riipu jakauman normaalisuudesta.

  Kun todistimme suurten lukujen lakia, joka oli pelkkä raja-arvotulos todennäköisyydelle, riitti mainiosti näinkin karkea arvio – pääasia oli, että sillä oli haluttu raja-arvo

Kolikkoesimerkki jatkuu

- Lukumäärä (ja frekvenssi) asettuu enintään 2 hajonnan päähän odotusarvosta tn:llä 0.955
- Lukumäärän hajonta $D(S_n) = \sqrt{(npq)} = 0.5 \cdot \sqrt{n}$
- Frekvenssin hajonta $D(f_n) = \sqrt{(pq/n)} = 0.5 / \sqrt{n}$

n	lkm:n hajonta D(Sn)	frekvenssin hajonta D(fn)
100	5	0.05
10 000	50	0.005
1 000 000	500	0.0005

Jos p on tuntematon ja sitä yritetään estimoida frekvenssillä, tarkkuuden lisääminen **yhdellä desimaalilla** vaatii **100-kertaisen** määrän kokeita!

Jatkuvuuskorjaus

- Kokonaislukuarvoisella muuttujalla X tapahtumat

  $X = k$

  $X \in (k-½, k+½)$

  ovat identtiset.

  Jos X:n jakaumaa approksimoidaan normaalilla, niin tapahtumalle ($X=k$) saadaan tn=**0** !!!

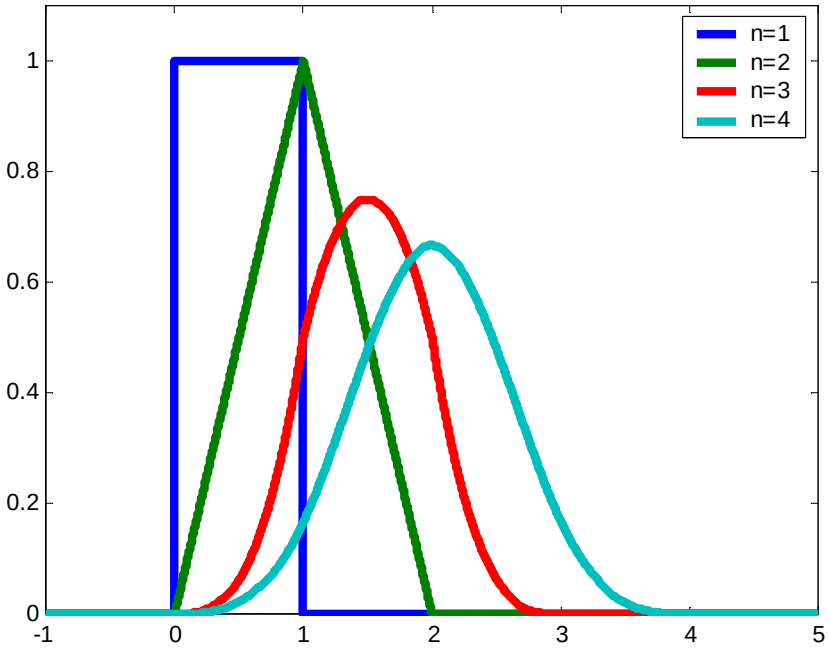
  Parempi tulos saadaan, kun kokonaislukua k edustamaan otetaan koko **väli** ($k-½, k+½$).

Kolikkoesimerkki: Vinon jakauman häntä

- Heitetään 30 kertaa painotettua kolikkoa, kruunan todennäköisyys p=0.1
- Tn. että saadaan enintään 2 kruunaa?

Normaaliapproksimaatio $S \sim N(3, 1.643^2)$	$\Phi((2-3) / 1.643)$	0.271
Normaali jatk.korjauksella $S \sim N(3, 1.643^2)$	$\Phi((2.5-3) / 1.643)$	0.380
$S \sim Poisson(3)$ Ks. Tuominen s. 53-54	f(0)+f(1)+f(2)	0.423
Tarkka $S \sim Bin(30, 0.1)$	f(0)+f(1)+f(2)	0.411

Tas(0,1)-muuttujien summan tarkka tiheysfunktio



$$f_{S_2}(x) = \begin{cases} x, & \text{kun } 0 < x \le 1 \\ 2 - x & \text{kun } 1 < x \le 2 \end{cases}$$

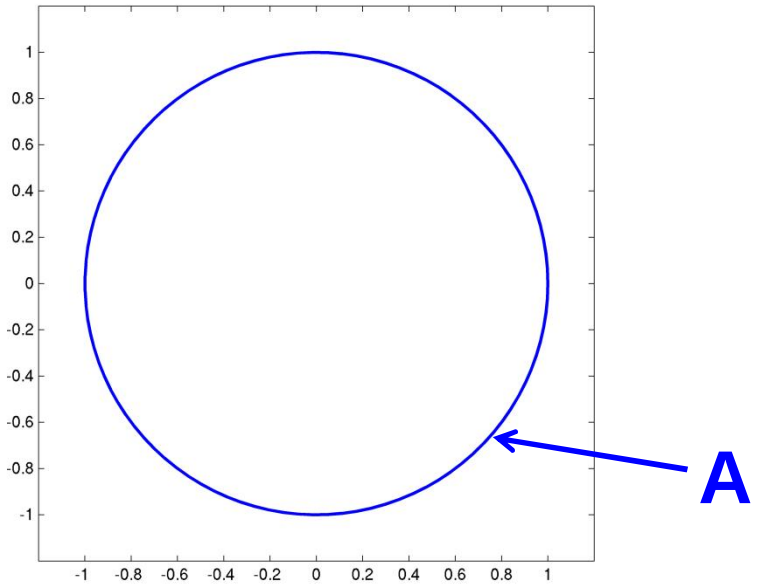
$$f_{S_3}(x) = \begin{cases} \frac{1}{2}x^2 & \text{kun } 0 < x \le 1 \\ -x^2 + 3x - \frac{3}{2} & \text{kun } 1 < x \le 2 \\ \frac{1}{2}(3-x)^2 & \text{kun } 2 < x \le 3 \end{cases}$$

Jne: jokaisella kahden kokonaisluvun välillä eri polynomi.
Ei kovin käytännöllistä suurilla *n*.

BERNOULLIN LAUSEEN SOVELLUTUS: MC-INTEGROINTI

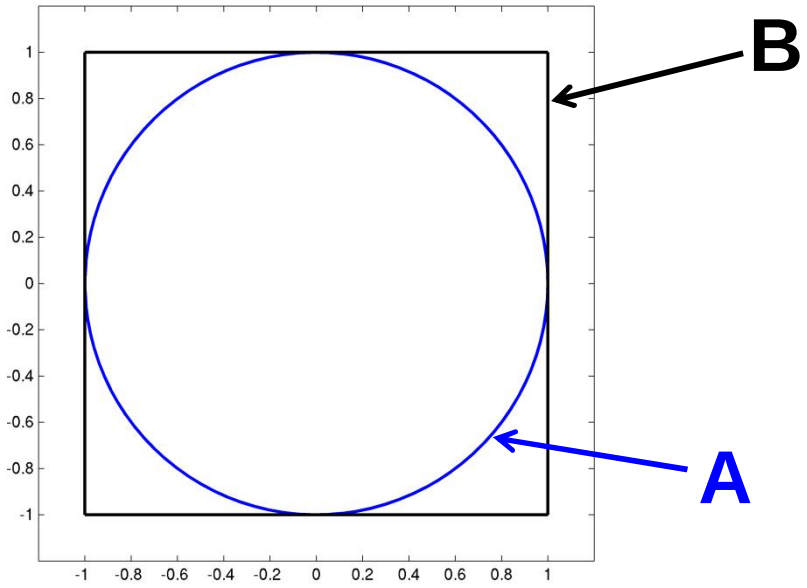
Tehtävässä ei todennäköisyyttä

- Mikä on mutkikkaan tasokuvion **A** pinta-ala?
  Osaamme vain testata, onko jokin piste sisällä vai ulkona:
  onko **sqrt($x^2 + y^2$) < 1**



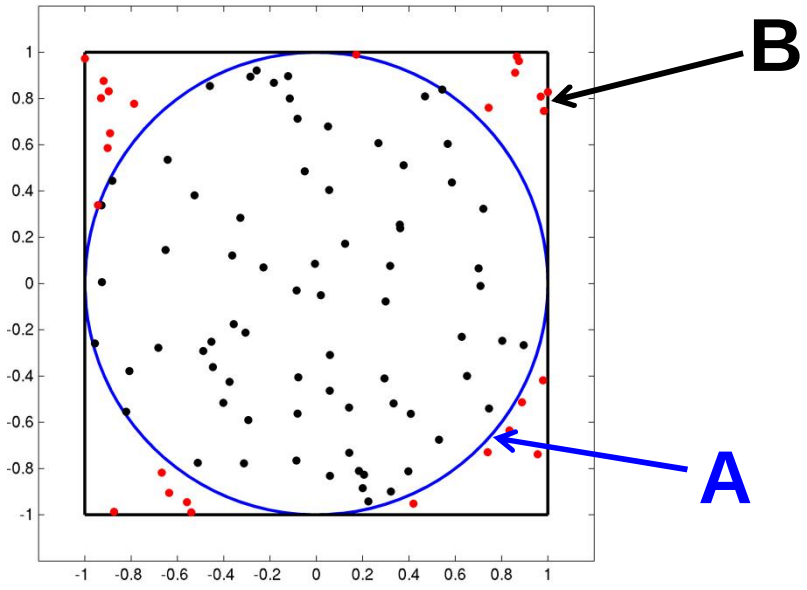
Muutetaan tehtävää

- Esimerkki: Mikä on mutkikkaan tasokuvion **A** pinta-ala?
  Osaamme vain testata, onko jokin piste sisällä vai ulkona.
- Ratkaisu: Piirrämme kuvion ympärille isomman (**B**),
  - jonka pinta-alan (= 4) tunnemme



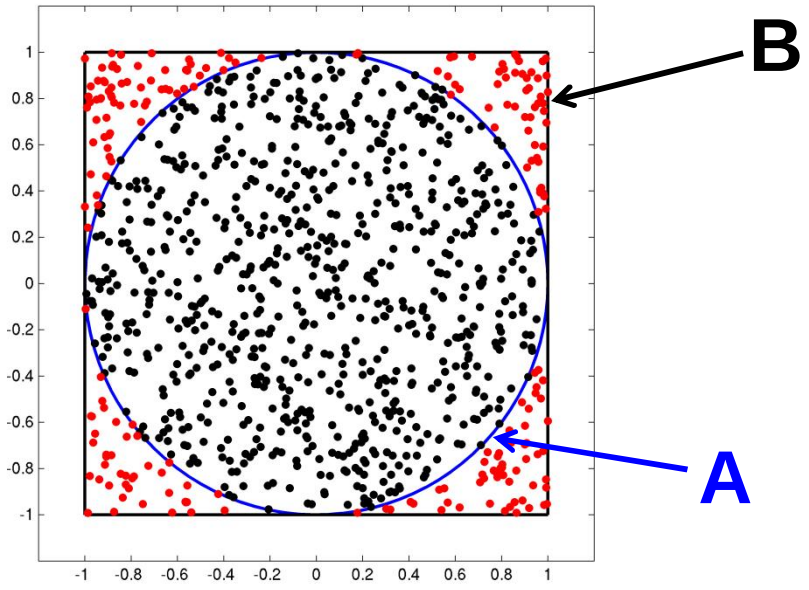
Muutetaan tehtävää

- Esimerkki: Mikä on mutkikkaan tasokuvion **A** pinta-ala?
  Osaamme vain testata, onko jokin piste sisällä vai ulkona.
- Ratkaisu: Piirrämme kuvion ympärille isomman (**B**),
  - jonka pinta-alan (= 4) tunnemme ja
  - josta osaamme arpoa (simuloida) pisteitä



Monte Carlo -integrointi

- Piste osuu kuvioon tn:llä $p = m(A) / m(B)$, m = pinta-ala
- n-kertainen toistokoe
- Bernoullin lause: osuus $f_n \approx p$
- Arvioimme, että $\boldsymbol{m(A) = p\, m(B) \approx f_n\, m(B)}$



Monte Carlo -integrointi

n	pisteitä B:ssä	$m(B) \approx$
100	80	3.200000
1 000	783	3.132000
10 000	7 849	3.139600
100 000	78 544	3.141760
1 000 000	785 132	3.140528

Samaa menetelmää voi periaatteessa soveltaa mielivaltaisessa *n*-ulotteisessa avaruudessa, esim. mikä on *n*-ulotteisen pallon tilavuus?

Keskeisen raja-arvolauseen perusteella voidaan arvioida integraalin likiarvon tarkkuutta (tunnetaan odotusarvo ja varianssi, approksimoidaan normaalijakautuneeksi). Yleisesti ottaen 100-kertaisella pistemäärällä saadaan yksi desimaali lisää tarkkuutta