Johdatus todennäköisyyslaskentaan
Kevät 2015

Luento 13 / 13

Jukka Kohonen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

Jakauman leveys

- Usein halutaan tunnusluku, joka kertoo miten tiiviisti mahdolliset arvot (eli X:n jakauma) **keskittyvät** odotusarvon lähelle
  - esim. ”koko arvojoukon leveys” ei oikein kerro tätä asiaa (vrt. tasa- ja kolmiojakauma)
- Eräs ratkaisu: **varianssi** (Tuominen s. 82)

  $\mathbf{Var(X) = E((X - \mu)^2)}$

  ja varianssin neliöjuuri eli **hajonta**
  - varianssilla paljon käteviä **laskusääntöjä**
  - varianssi myös **tunnetaan** useille tutuille jakaumille
- Varianssille käytetään vaihtelevasti merkintöjä Var(X) ja $D^2(X)$, ne tarkoittavat samaa.

Varianssin idea

- Oletetaan tunnetuksi $E(X) = \mu$.
- Tiedetään, ettei X aina (eikä edes kovin usein) osu odotusarvoonsa. (Ehkä ei koskaan)
- Katsotaan, paljonko se meni huti (= **poikkeama**) ja korotetaan poikkeama toiseen (= **neliöpoikkeama**)
- Jokaiseen X:n mahdolliseen arvoon liittyy vastaava neliöpoikkeama $(X-\mu)^2$
- X:n mahdollisilla arvoilla on todennäköisyydet, joten voidaan laskea, miten neliöpoikkeama on jakautunut. Lasketaan sen odotusarvo ja nimitetään sitä varianssiksi.
- Suuri varianssi siis merkitsee, että X usein poikkeaa paljon odotusarvostaan
- Pienimillään varianssi voisi olla 0, jos X on aina = μ

Summan varianssi, riippumattomat muuttujat

Olkoon S = X + Y.

Osataan tietysti laskea E(S) = E(X) + E(Y)

Lasketaan auki Var(S).

Kaava sievenee kummasti, jos pystytään käyttämään riippumattomien muuttujien tulokaavaa E(XY) = E(X) E(Y)

→Lause 3.2.5(iii) (sivu 83)

Sovellutus: Binomijakauman varianssi ***npq***

Helppoja hajonnan ominaisuuksia

Tuominen s. 83: Lause 3.2.5

- Eräissä jakauman muunnoksissa on helppo(?) päätellä, mitä hajonnalle tapahtuu

- Vakion **lisääminen**: Hajonta ei muutu
- Vakiolla **kertominen**: Hajonnalle sama kerroin (tai itseisarvo, jos vakio on negatiivinen)

- Kahden **riippumattoman** sm:n **summa**: **varianssit** voi laskea yhteen
  (hajontoja ei voi laskea yhteen – hajonta on varianssin neliöjuuri)

Eräiden jakaumien tunnuslukuja

Jakauma	Odotusarvo	Hajonta
Tas(0, 1)	1/2	$1 / \sqrt{12}$
Tas(a, b)	(a+b) / 2	($b-a$) / $\sqrt{12}$
N(0, 1)	0	1
N(μ, σ^2)	μ	σ
Exp(1)	1	1
Exp(λ)	$1 / \lambda$	$1 / \lambda$

Näissä kaikissa ”yleinen jakauma” saadaan ”perusjakaumasta” jollain venytyksellä (vakiokerroin) ja siirrolla (vakiolisäys), ja tämä näkyy tunnusluvuissakin edellä opitulla tavalla.

Hajonta on yleensä helpommin ymmärrettävä tunnusluku kuin varianssi (varianssia käytetään lähinnä eräiden laskukaavojen, kuten summakaavan takia)

Vakiokerroin ja vakiolisäys

- Puhelinpalvelun jonotusaika minuutteina X ~ Tas(0, 10)

  E(X) = (0+10)/2 = 5.00 min

  D(X) = 10 · √(1/12) = 2.89 min

  D(X) voidaan esim. laskea integroimalla; tai muistaa kaava Tas(a,b):lle; tai päätellä Tas(0,1)-jakauman hajonnasta, joka on $1/\sqrt{12} \approx 0.289$

- **Vakiokerroin**: jonotus maksaa 0.20 €/min, hinta Y = 0.2·X

  E(Y) = 0.2 · E(X) = 0.2 · 5 = 1.00 € sama kerroin

  D(Y) = 0.2 · D(X) = 0.2 · 2.89 = 0.58 € sama kerroin

- **Vakiolisäys**: puheluaika 3 min, kokonaisaika Z = X+3

  E(Z) = E(X) + 3 = 5 + 3 = 8.00 min sama lisäys

  D(Z) = D(X) = 2.89 min hajonta ei kasva!

Odotusarvon ja varianssin kaavoja

Muunnos	**Odotusarvo** (Lause 3.1.1)	**Varianssi** (Lause 3.2.5)
vakion lisäys	$E(X + b) = E(X) + b$	$Var(X + b) = Var(X)$
vakiokerroin	$E(aX) = a \cdot E(X)$	$Var(aX) = a^2 \cdot Var(X)$
sm:ien summa	$E(X+Y) = E(X) + E(Y)$	$Var(X+Y) = Var(X) + Var(Y)$ *jos* $X \perp\!\!\!\perp Y$
sm:ien tulo	$E(XY) = E(X) \cdot E(Y)$ *jos* $X \perp\!\!\!\perp Y$	
yleinen muunnos $g(X)$	$\mathrm{E}[g(X)] = \int g(x) f(x) dx$ (Lause 3.1.8)	

SUURTEN LUKUJEN LAKI

Mihin pyritään

- Olemme empiirisesti huomanneet, että otoksen histogrammi muistuttaa jakauman tiheysfunktiota.

- Voisiko otoksen avulla arvioida myös tunnuslukuja, esim. jakauman odotusarvoa?
- Ilmeinen ehdokas olisi otoksen aritmeettinen keskiarvo. MATLAB: **`sum(x)/n`** tai **`mean(x)`**

- Todistetaan lause, jonka mukaan otoskeskiarvo on ”pitkän päälle” todennäköisesti hyvä arvio odotusarvolle.
  - Yritetään tehdä tämä yleisesti (monille eri jakaumille)
  - Ajatellaan ensin erikoistapausta: Bernoullin lausetta

Kohti Bernoullin lausetta

Tuominen s. 93

Toistokoe, ***n*** toistoa tn:llä ***p***

Z_n = onnistumisten lukumäärä

f_n = onnistumisten osuus = Z_n / n

Tiedetään vanhastaan, että: $Z_n \sim \mathbf{Bin}(n, p)$

Siispä onnistumisten määrän odotusarvo on ***np*** ja osuuden odotusarvo ***p***.

Myös jakaumien moodit ovat suunnilleen tässä kohdassa (Tuominen s. 51)

Mutta **täsmälleen** odotusarvoon ei osuta helposti (harj. 6:8)

Osumisen tn jopa **pienenee**, kun *n* kasvaa. Esim. kun *p*=0.3:

n	np	$P(Z_n = np) \approx$
10	3	0.27
100	30	0.087
1 000	300	0.028
1 000 000	300 000	0.000 87

Osutaanko edes lähelle?

Z_n = onnistumisten lukumäärä (binomijakautunut)

f_n = onnistumisten osuus = $Z_n \,/\, n$

ε = osuuden tarkkuusvaatimus (voimme valita tämän vapaasti)

Sanomme, että f_n osuu lähelle, kun $|\, f_n - p \,| < \varepsilon$

Lähelle osumisen tn **kasvaa**, kun *n* kasvaa – ainakin siltä näyttää:
Esim. kun p=0.3 ja **ε=0.01**, lähelle osutaan kun $0.29 < f_n < 0.31$:

n	f_n *oltava välillä*	Z_n *oltava välillä*	P(lähellä) ≈
100	(0.29, 0.31)	{30}	0.087
1 000	(0.29, 0.31)	(290, 310)	0.488
10 000	(0.29, 0.31)	(2 900, 3 100)	0.970
100 000	(0.29, 0.31)	(29 000, 31 000)	0.999999999946

P(lähellä) = summa binomijakauman pistetodennäköisyyksistä

Bernoullin lause (Tuominen 3.5.4, s. 93)

- Valitaan mikä tahansa tarkkuusvaatimus **ε** > 0
- Bernoullin lause kertoo **raja-arvotuloksen:**

$$P\left(\left|f_n - p\right| < \varepsilon\right) \to 1$$

kun $n \to \infty$.

- Tämän todistaminen edellyttää, että pystytään sanomaan jotain binomijakauman todennäköisyydestä **isolla välillä**. Suoraviivainen lasku, **ptnf yhteenlasku** ei ole helppoa, kun termien määräkin kasvaa *n*:n mukana
- Tunnemme binomijakauman **varianssin** *npq*, mutta onko siitäkään apua? Miten hajonta liittyy todennäköisyyksiin?
- Käytetään yleistä matemaattista menetelmää: todistetaan joku **epäyhtälö**, kun ei osata todistaa tarkkaa yhtälöä
- Todistus perustuu kahteen epäyhtälöön: Markovin ja Tsebysevin. Tutustumme niihin seuraavaksi.

EPÄYHTÄLÖITÄ POIKKEAMIEN ARVIOIMISEEN

Epäyhtälöitä

- Jos jakaumaa ei tunneta tarkasti, mutta tunnetaan $E(X)$ ja ehkä $Var(X)$,
  niin **suurten poikkeamien todennäköisyyksiä**
  (jakauman ”**häntiä**”) voidaan arvioida erilaisilla epäyhtälöillä.

- Arviot ovat kuitenkin aika karkeita.

Markovin epäyhtälö (Tuominen s. 81)

- Jos varmasti **X ≥ 0**, ja **E(X) tunnetaan**,
  voidaan arvioida häntätodennäköisyyttä mistä tahansa kohdasta ***a*** oikealle (äärettömiin asti)

$$P(X \geq a) \leq E(X) / a$$

Jakaumasta ei tarvitse tietää muuta kuin em. asiat.

Todistusidea:

Odotusarvo on summa tai integraali.
Rajoitetaan sen termejä tai integroitavaa **alhaalta**
→ saadaan odotusarvolle **alaraja**
→ saadaan todennäköisyydelle **yläraja**.

Summan tai integraalin rajoittaminen

- Yleisiä matemaattisia havaintoja:

  – Jos termeittäin pätee $a_i \leq b_i, \quad \forall i,$
  pätee myös summille $\sum a_i \leq \sum b_i$

  – Jos pätee $f(x) \leq g(x), \quad \forall x,$
  pätee myös $\int f(x)dx \leq \int g(x)dx$

Markovin epäyhtälö, diskreetti X

Oletetaan yksinkertaisuuden vuoksi, että X on kokonaislukuarvoinen

- **Tiedetään, että $X \geq 0$**, ja **$E(X)$ tunnetaan.**
- Merkitään
  pistetodennäköisyydet $p_k = P(X = k)$
  häntätodennäköisyys $q_a = P(X \geq a) = p_a + p_{a+1} + \ldots$

$$E(X) = \sum_{k=0}^{\infty} k\, p_k$$

Odotusarvo esitetty summana

$$= \left(\sum_{k=0}^{a-1} k\, p_k \right) + \left(\sum_{k=a}^{\infty} k\, p_k \right)$$

Summa hajotettu kahteen osaan

$$\geq \left(\sum_{k=0}^{\infty} 0\, p_k \right) + \left(\sum_{k=a}^{\infty} a\, p_k \right)$$

Kaikkia termejä **rajoitettu alhaalta**.
Ensimmäinen osasumma häviää.
Toisessa otetaan a ulos summasta

$$= 0 + a \cdot q_a$$

$$= a \cdot q_a,$$

joten puolestaan $q_a \leq E(X)/a$, eli Markovin epäyhtälö pätee.

Markovin epäyhtälö, jatkuva *X*

- Todistus samaan tapaan kuin diskreetillä *X*:llä. Odotusarvo on integraali, jota rajoitetaan alhaalta, nytkin saadaan

$$E(X) \geq pa,$$

josta seuraa

$$p \leq E(x)/a.$$

Tšebyševin epäyhtälö (Tuominen s. 85)

- Jos ***μ* = E(*X*) ja *σ* = D(*X*) molemmat tunnetaan**, voidaan häntä-tn arvioida tehokkaammin:

$$P\left(\left|X-\mu\right| \geq k\sigma\right) \leq \frac{1}{k^2}$$

- Siis tn, että *X* poikkeaa odotusarvostaan ”yli *k*:n hajonnan verran”, on enintään $1/k^2$

- *X* ei tarvitse olla epänegatiivinen (kuten Markovissa)
- Tiukempi raja, koska nimittäjässä on toinen potenssi

- Todistuksessa Markov-epäyhtälöä sovelletaan neliöpoikkeamiin $(X-\mu)^2$, niiden odotusarvohan on = Var(X)

Esim: Exp-jakauman häntä-tn eri tavoilla

- Koneen elinikä $X \sim \text{Exp}(0.1)$. Arvioidaan todennäköisyyttä $P(X \geq a)$ a:n eri arvoilla ja eri menetelmillä.
- $E(X)=10$, $D(X)=10$, kertymäfunktio tunnetaankin $F(x)=1-\exp(-0.1x)$

	Markov	Tsebysev	Tarkasti kf:stä
$P(X \geq 10)$	< 1		0.368
$P(X \geq 20)$	$< 1/2$	< 1	0.135
$P(X \geq 30)$	$< 1/3$	$< 1/4 = 0.25$	0.050
$P(X \geq 40)$	$< 1/4$	$< 1/9 = 0.11$	0.018
$P(X \geq 200)$	$< 1/20$	$< 1/361 = 0.0028$	$2.1 \cdot 10^{-9}$

- Suurille poikkeamille Tšebyšev on tarkempi kuin Markov.
- Vielä paljon tarkempi on tarkka kertymäfunktio F, jos se tunnetaan!
- Mutta epäyhtälöillä saatetaan saada johonkin tarkoitukseen ”riittävä” arvio.
- Tšebyševistä on apua esim. suurten lukujen lain todistamisessa.

Tšebyšev → Suurten lukujen laki

Tuominen s. 91

- $(X_1, X_2, ...)$ jono riippumattomia sm:ia, joilla sama odotusarvo $\boldsymbol{\mu}$ ja varianssi $\boldsymbol{\sigma^2}$
- valitaan mv. tarkkuusvaatimus $\boldsymbol{\varepsilon} > 0$

Osasumma S_n $= X_1+...+X_n$

- Osasumman odotusarvo $= \boldsymbol{n\mu}$ miksi?
- Osasumman varianssi $= \boldsymbol{n\sigma^2}$ miksi?

Keskiarvo $= (X_1+...+X_n) / n$

- Keskiarvon odotusarvo $= \boldsymbol{\mu}$ miksi?
- Keskiarvon varianssi $= \boldsymbol{\sigma^2 / n}$ miksi?

Tšebyšev → Suurten lukujen laki

Tuominen s. 91

valitaan mv. tarkkuusvaatimus $\varepsilon > 0$

Otoskeskiarvo $= (X_1 + ... + X_n) / n$

- Otoskeskiarvon odotusarvo $= \mu$
- Otoskeskiarvon varianssi $= \sigma^2 / n$
- Otoskeskiarvon hajonta $= \sigma / \mathrm{sqrt}(n)$ miksi?

Otoskeskiarvon **hajonta pienenee** *n*:n kasvaessa.

Tällöin kiinteä tarkkuusvaatimus **ε** merkitsee yhä suurempaa **kerrointa (*k*)** hajonnalle, jolloin Tsebysevin epäyhtälön perusteella **tn niin suureen poikkeamaan** odotusarvosta **pienenee**.

On siis yhä todennäköisempää (tn → 1), että otoskeskiarvo osuu alle **ε**:n päähän odotusarvostaan eli ***μ***:stä. = Suurten lukujen laki

SLL erikoistapaus: Bernoullin lause

- $(X_1, X_2, ...)$ jono riippumattomia **indikaattorimuuttujia** toistokokeelle
- indikaattorien osasumma = onnistumisten **lukumäärä**
- lukumäärä on tunnetusti binomijakautunut,
  $E(S_n) = np$
  $Var(S_n) = npq$

- Indikaattorien keskiarvo = onnistumisten **osuus**
- SLL näille indikaattoreille → tn, että onnistumisosuus on ”lähellä” odotusarvoaan, lähenee rajatta 1:tä
- Bernoullin lause kertoo, miksi **todennäköisyydellä** on jotain tekemistä toistokokeessa toteutuvan **osuuden** kanssa.

Varianssin ominaisuuksia

- Jos satunnaismuuttujan *X* varianssi tunnetaan,
  - Mikä on Var(Y), jos Y = aX ?
  - Mikä on Var(Y), jos Y = X + b ?

SUMMAN $Y = X_1 + \ldots + X_n$ JAKAUMA

2.3 Satunnaismuuttujien summa

Herra K:n bussimatka koostuu osista

- bussin odotus Y ~ Tas(0, 4) E(Y)=2 min
- ajoaika Z ~ Tas(15, 25) E(Z)=20 min

Koko matka-aika **M** = Y+Z

- vähintään 0 + 15 = 15 min
- enintään 4 + 25 = 29 min

- odotusarvo E(M) = E(Y) + E(Z) = 22 min

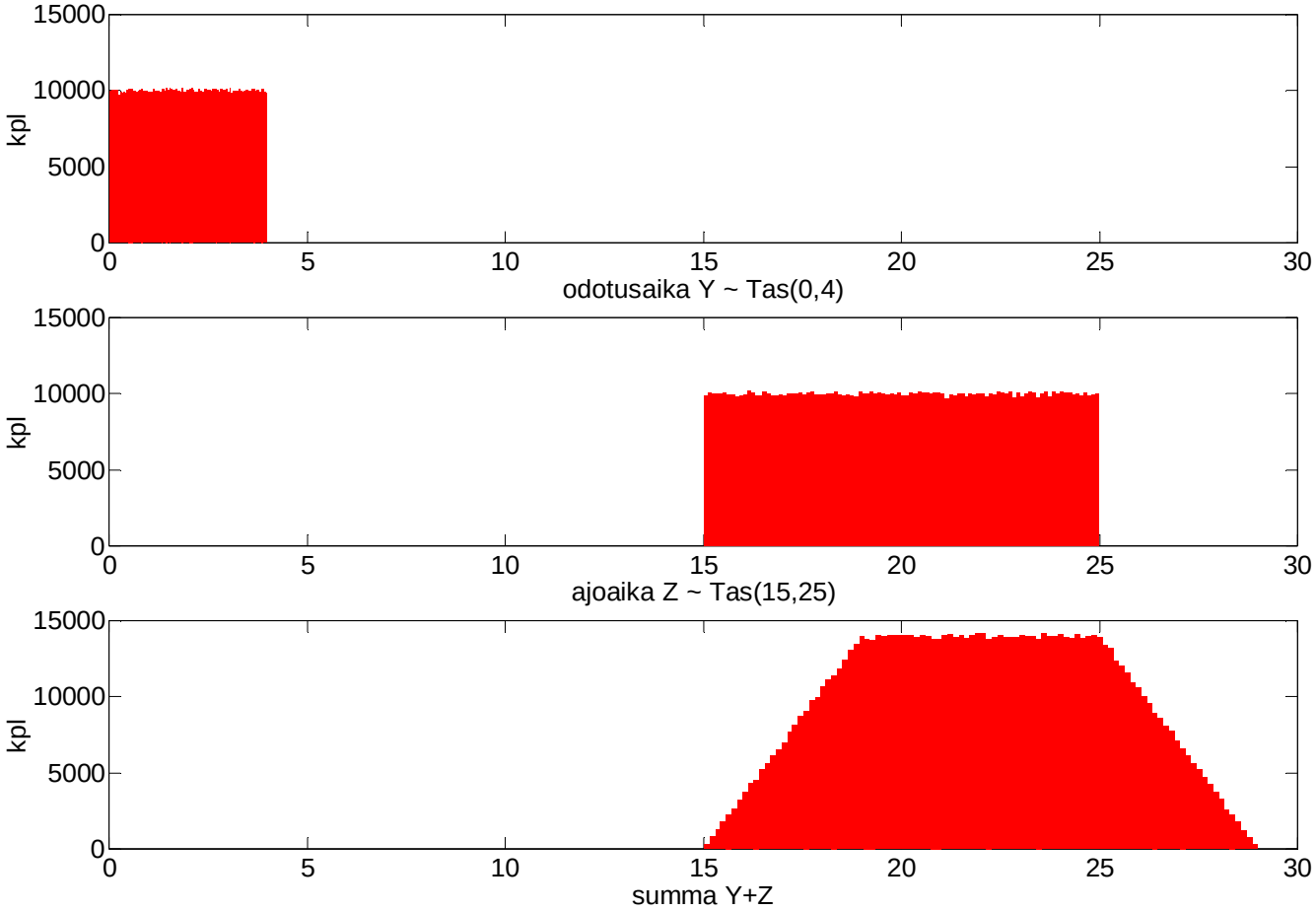
Mikä on M:n jakauma välillä (15, 29)?

- Odotusarvo ei kerro kaikkea jakaumasta.
  Jakauma voi esim. olla hyvin keskittynyt odotusarvon lähelle, tai sitten ei

Empiirinen kokeilu

Bussin odotusaika Tas(0,4) ja ajoaika Tas(15,25). Miten summa on jakautunut?

Arvotaan miljoona kertaa Y~Tas(0,4) ja Z~Tas(15,25) ja lasketaan summat.
Empiirinen histogrammi antaa tällöin hyvän käsityksen tiheysfunktion muodosta.
Esim. Matlabissa Y = unifrnd(0,4,1,1e6); Z=unifrnd(15,25,1,1e6); M=Y+Z; hist(M)



Y ~ Tas(0,4)

Z ~ Tas(15,25)

M = Y+Z:
ei tasajakauma
(Miten poikkeaa?
Miksi?)

Jos halutaan laskea tarkasti

- Summan M=X+Y jakauma voidaan laskea ns. **konvoluutiolla** (Tuominen 71-72)
- Voidaan myös **järkeillä geometrisesti**, miten X ja Y voivat sijoittua: ajatellaan, että piste (X,Y) on tasajakautunut suorakulmiossa (0, 4) × (15, 25)
- Tähän tarvitaan 2-ulotteisen jakauman käsite. Vrt. Tuominen s. 131 esimerkki 5.3.5: Tasainen jakauma alueessa *A*

Jos halutaan approksimoida

- Summan voidaan **arvioida** olevan **suunnilleen** **normaalijakautunut**,
  ja lasketaan sen parametrit μ ja σ
- Tähän tarvitaan
  - normaalijakauman käsite
  - varianssin käsite

Esimerkki: **Neljän** tasajakautuneen summa

(Herra K matkustaa kahdella bussilla peräkkäin)

odotus1	***X***	~ Tas(0, 4)	E(*X*) = 2
ajoaika1	***Y***	~ Tas(10, 14)	E(*Y*) = 12
odotus2	***Z***	~ Tas(0, 6)	E(*Z*) = 3
ajoaika2	***W***	~ Tas(20, 26)	E(*W*) = 23
Matka-aika	***M***	= *X*+*Y*+*Z*+*W*	E(*M*) = 40

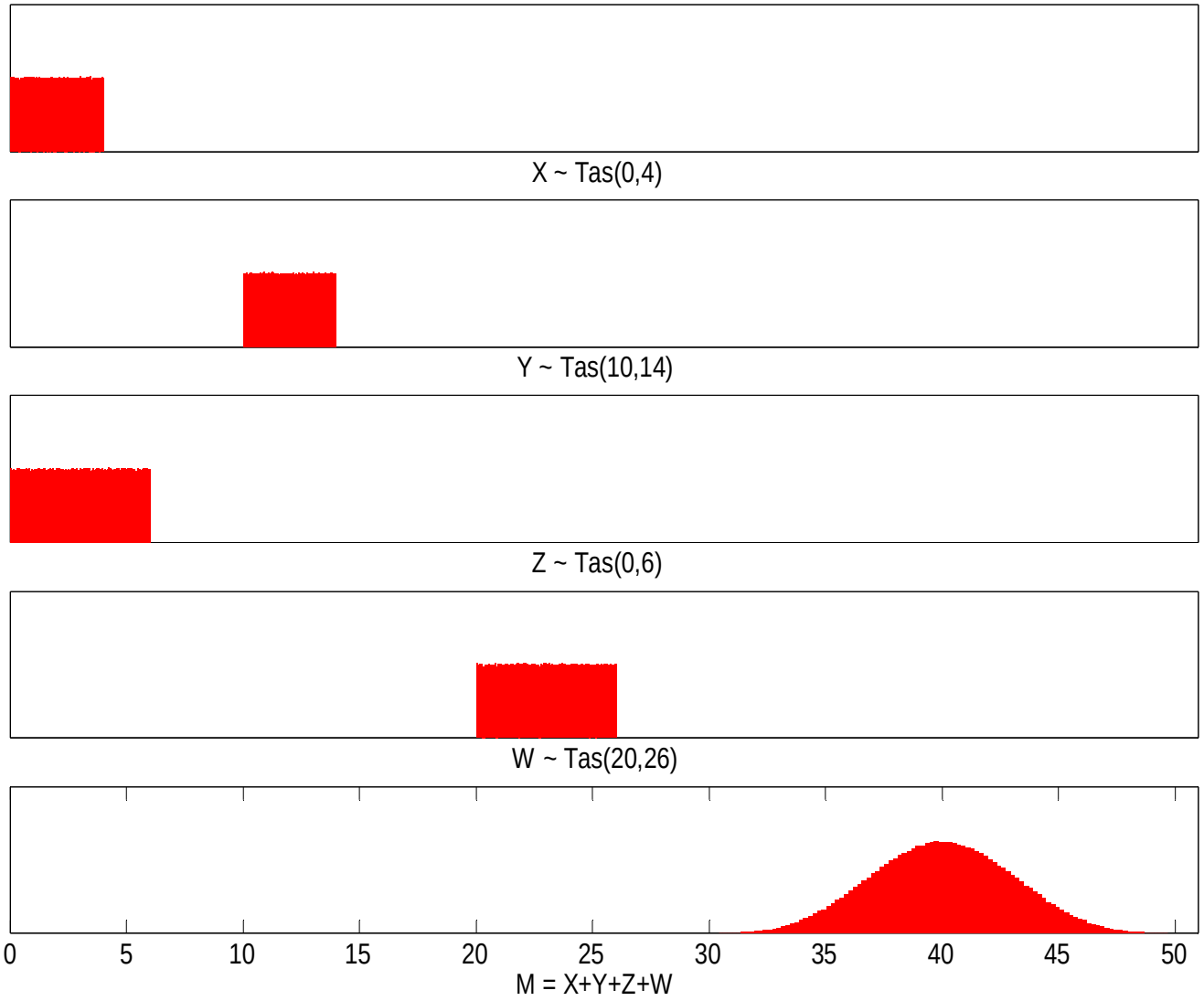
- *M* vähintään 0+10+0+20 = 30
- *M* enintään 4+14+6+26 = 50

Minkä muotoinen jakauma *M*:llä on?

Ei varmaan tasajakauma, vaikka onkin välillä (30, 50) ja vaikka odotusarvo 40 on välin keskellä.

Esimerkki: **Neljän** tasajakautuneen summa

(Herra K matkustaa matkustaa kahdella bussilla)



Tutun näköinen jakauma?

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

Jakaumien yhteenlaskuominaisuuksia

Kun X $\perp\!\!\!\perp$ Y, samasta jakaumasta, ja S=X+Y, niin summan jakauma usein tunnetaan:

Diskreettejä yhteenlaskettavia:

- X,Y ~ kolikko → S ~ binomijakauma
- X,Y ~ noppa → S ~ diskreetti kolmion muotoinen
- X,Y ~ Geom → S ~ negatiivinen binomijakauma

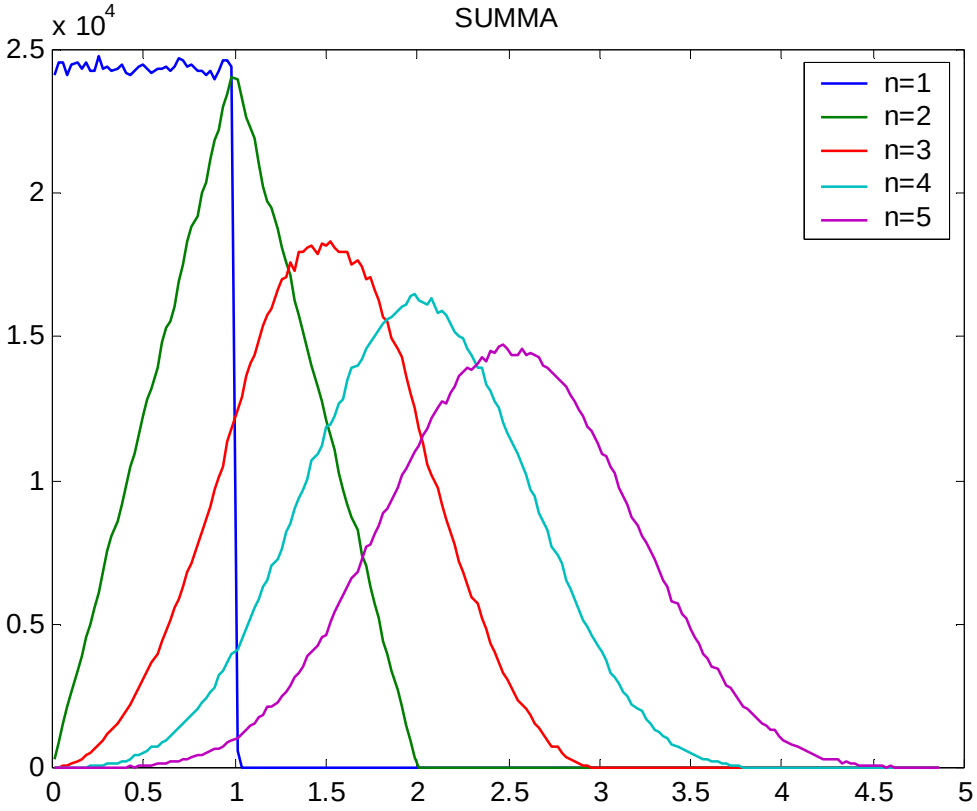
Jatkuvia yhteenlaskettavia:

- X,Y ~ Tas → S ~ jatkuva kolmiojakauma
- X,Y ~ Exp → S ~ gammajakauma

- Summien **täsmälliset** jakaumat ovat **erilaisia**.
- Kuitenkin jos yhteenlaskettavia on monta, nämä erilaiset ”summajakaumat” **muistuttavat** toisiaan.

KESKEINEN RAJA-ARVOLAUSE

Tas(0,1)-muuttujien summa



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

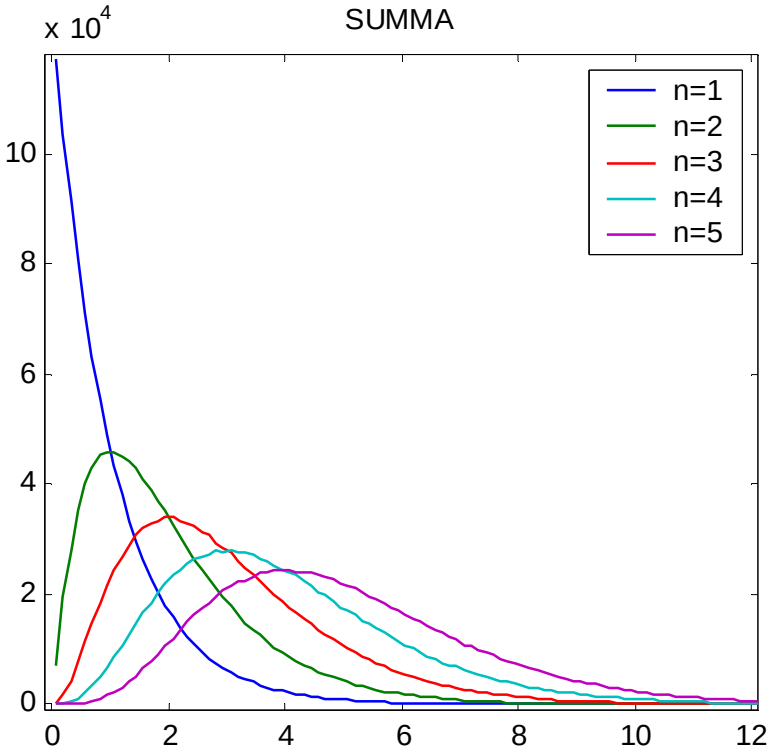
Summa
$\mathbf{S_n = \Sigma X_i}$

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

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

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)

$X+Y \sim N($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** !!!

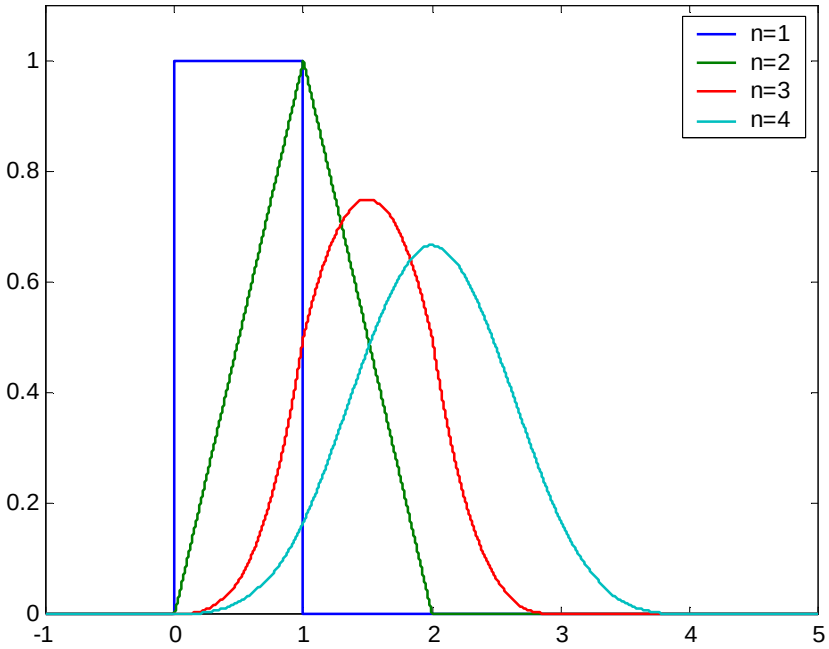
  Parempi tulos saadaan, kun kokonaislukua k edustamaan otetaan koko **väli** ($k-½, k+½$).

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



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

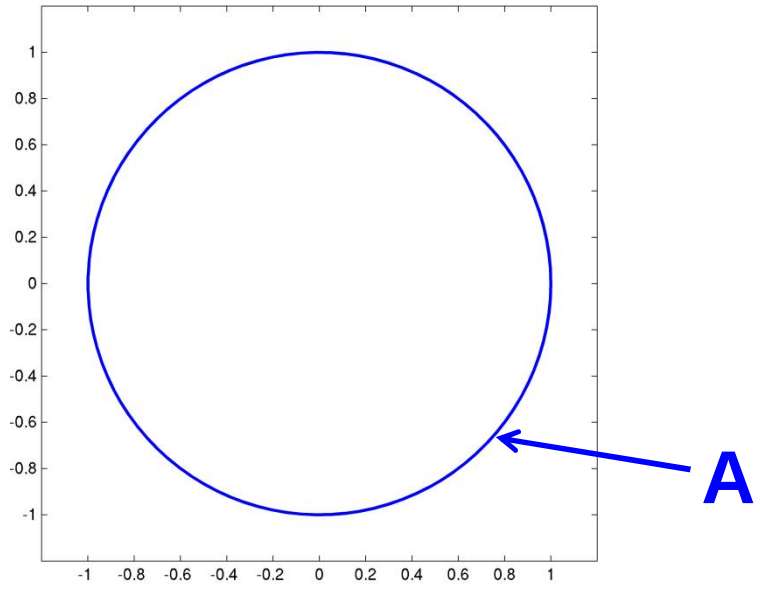
$$f_{S_3}(x) = \begin{cases} \frac{1}{2}x^2 & \text{kun } 0 < x \le 1 \\ -x^2 + 3x - \frac{3}{2} & \text{kun } 1 < x \le 2 \\ \frac{1}{2}(3-x)^2 & \text{kun } 2 < x \le 3 \end{cases}$$

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

BERNOULLIN LAUSEEN SOVELLUTUS: MC-INTEGROINTI

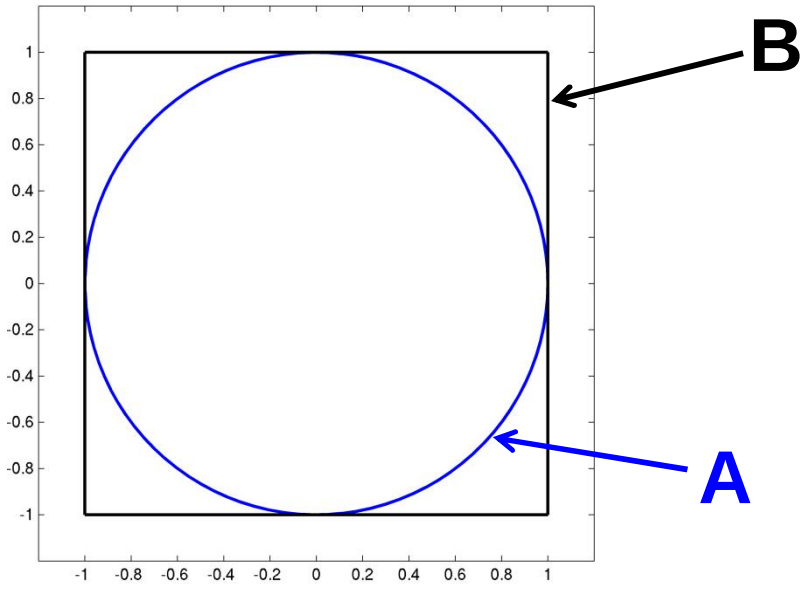
Tehtävässä ei todennäköisyyttä

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



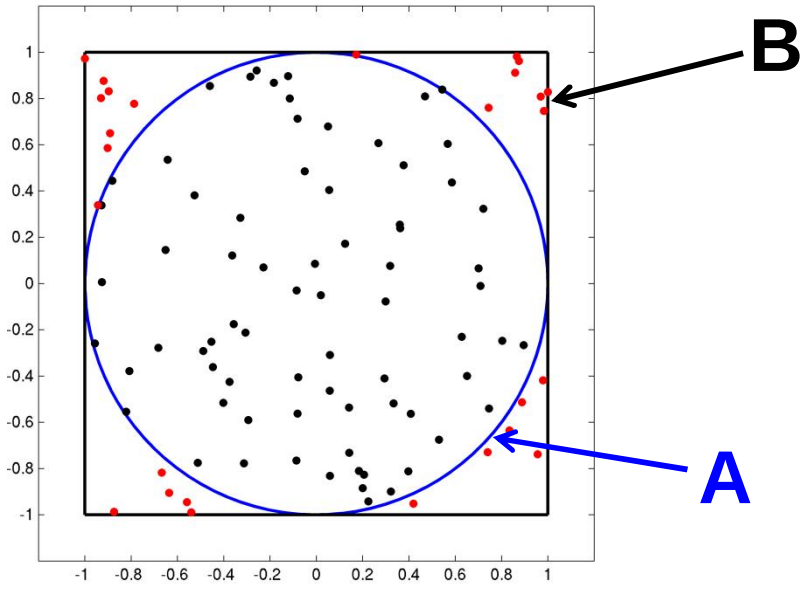
Muutetaan tehtävää

- 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



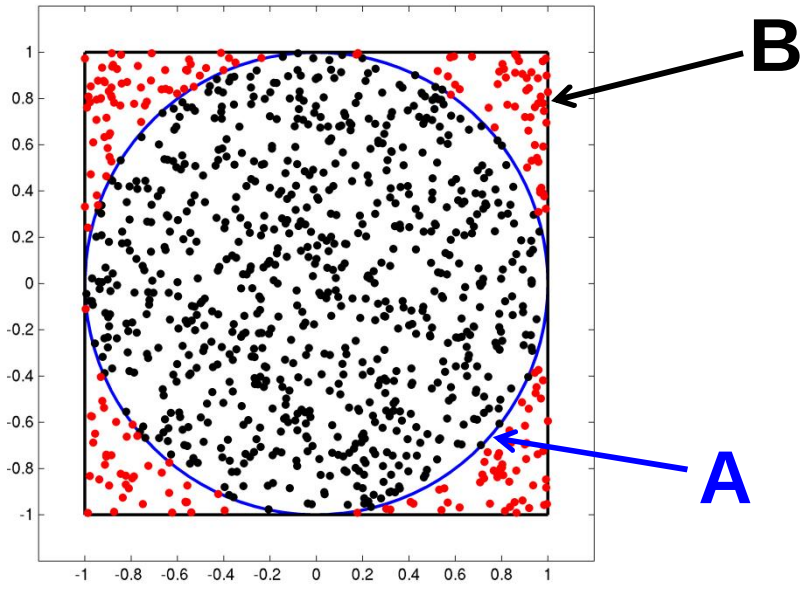
Muutetaan tehtävää

- 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.