Johdatus todennäköisyyslaskentaan
Kevät 2014

Luento 9 / 13

Jukka Kohonen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

Geom. jakauman sovellutus

- Harjoitustehtävä 4:14 (herrojen noppapeli)

- Pelin kestoa voidaan mallintaa geometrisella jakaumalla.
- A voittaa joss peli päättyy parittoman numeroisella heitolla, ts. epäonnistuneita heittoja on parillinen määrä
- Ts. A voittaa $\Leftrightarrow$ X on parillinen, missä $X \sim \text{Geom}(1/6)$

Binomijakauman odotusarvo

- ***n***-kertainen toistokoe tn:llä ***p***
- indikaattorimuuttujat ilmaisevat kokeiden onnistumista (0 tai 1)
  $X_1, ..., X_n$ Jokaisen odotusarvo = p.
- Onnistumisten määrä = indikaattorien summa
  $Y = X_1 + ... + X_n$
- Olemme määritelleet käsitteen ”binomijakauma”:
  $Y \sim \text{Bin}(n, p)$.
- Odotusarvon lineaarisuudesta seuraa
  $\text{E}(Y) = \text{E}(X_1) + ... + \text{E}(X_n)$
  $= p + ... + p$ (n kpl termejä)
  $= np$.

Esimerkki. Heitetään noppaa 100 kertaa ja kirjataan kuutosten määrä.

Määrän odotusarvo = $100 \cdot (1/6) \approx 16.667$

Geom. jakauman odotusarvo

- toistokoe tn:llä ***p***, toistetaan kunnes onnistuu, lasketaan epäonnistumiset = X

  $X \sim \text{Geom}(p)$

- Nyt E(X) voidaan laskea muutamallakin tavalla.
  - Tuominen s. 55 esittää ns. derivointikikan
  - Taululla: Hajotetaan odotusarvoa esittävä summa pienempiin osiin
  - Muitakin tapoja on, esim. eräiden indikaattorimuuttujien avulla

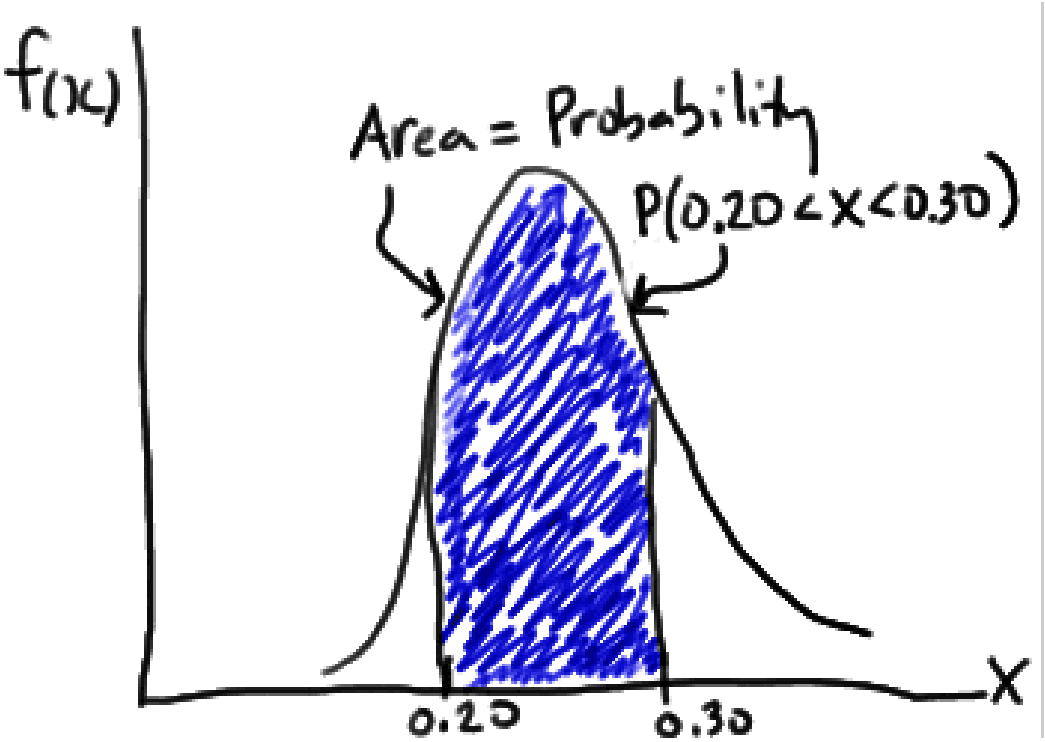
- Kaikilla menetelmillä sama tulos E(X) = q/p

Esimerkki. Heitetään noppaa kunnes saadaan vähintään viitonen.

$p = 2/6 = 1/3$

$q = 4/6 = 2/3$

Epäonnistumisten määrän odotusarvo = $(q / p) = 2$



JATKUVA SATUNNAISMUUTTUJA

Jatkuva tasajakauma

- Idea: Reaaliarvoinen sm saa jonkin arvon välillä (a,b)
- Kaikki yhtä pitkät välit ovat yhtä todennäköisiä
- Eripituisilla väleillä tn verrannollinen välin pituuteen – vrt. Jussin junamatkat
- Yhteys mittateoriaan

Tiheysfunktio

- Jos X:n tiheysfunktio on f,
  niin f(x) **ilmaisee verrannollisuuskertoimen**:
  tn osua (lyhyelle) välille (x:n lähellä) on välin pituus kertaa f(x).

- Huom: f(x) **ei ole todennäköisyys**, että X=x
- Esim. ennuste lämpötilasta, joka tasajakautunut välillä (20, 30)
- Koska f(x) ei ole todennäköisyys, se voi olla suurempikin kuin 1. (Mitä tämä tarkoittaa?)

Kertymäfunktio

- Vaikka tiheysfunktio antaa intuitiivisemman käsityksen jakaumasta (mihin jakauma keskittyy), kertymäfunktio on keskeinen laskennallinen työkalu jakaumien käsittelyssä.
- F(x) vastaa kysymykseen ”mikä on tn, että $X \leq x$”.
- Esim. lämpötilalle F(25) = ½ tarkoittaa, että on ½ todennäköisyys, että lämpötila on enintään 25 astetta.

- Kertymäfunktion on pakko olla kasvava – miksi?
- Kertymäfunktio kasvaa arvojoukossa vasemmalta 0:sta oikealle 1:een – miksi?
- Kertymäfunktio on yleensä tf:n integraali, ja tf on kertymäfunktion derivaatta – miksi?

Jatkuvan sm:n odotusarvo

- Odotusarvo määritellään melko analogisesti **integraalina** arvojoukon yli, integroitavana suureena on tiheysfunktio.
- Lineaarisuus yleensä pätee nytkin (Fubinin lause: summan ja integraalin vaihdannaisuus)

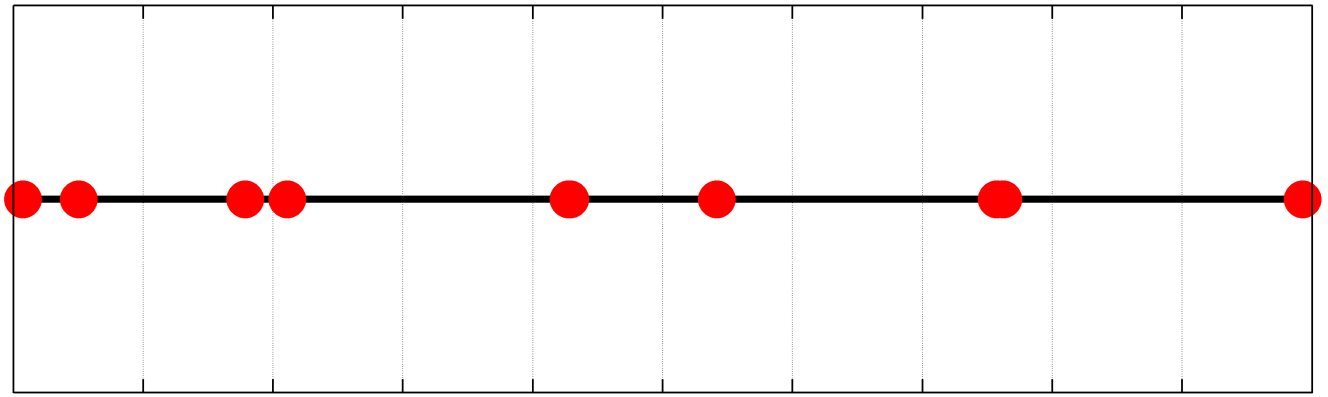
Epätasaisia jakaumia

- Eksponenttijakauma (Tuominen s. 59)
  muistuttaa geometrista jakaumaa, mutta onkin jatkuva

- Ts. ”yrityksiä” on jatkuvasti, ei vain kokonaislukujen kohdalla. Esim.
  - Alkeishiukkanen voi hajota milloin tahansa, mikä on hiukkasen elinikä?
  - Laite voi hajota milloin tahansa, miten kauan se toimii?
    (”Satunnainen” hajoaminen, ei kuluminen)
  - Tuulilasiin voi osua kivi milloin tahansa, miten kauan voidaan ajaa ennen kuin ensimmäinen kivi osuu?

- Huom: mediaani ei ole odotusarvon kohdalla
  (lasketaan kertymäfunktion avulla!)

  (mediaani = se piste, jonka molemmin puolin on puolet todennäköisyydestä, tarkemmin Tuominen s. 89)

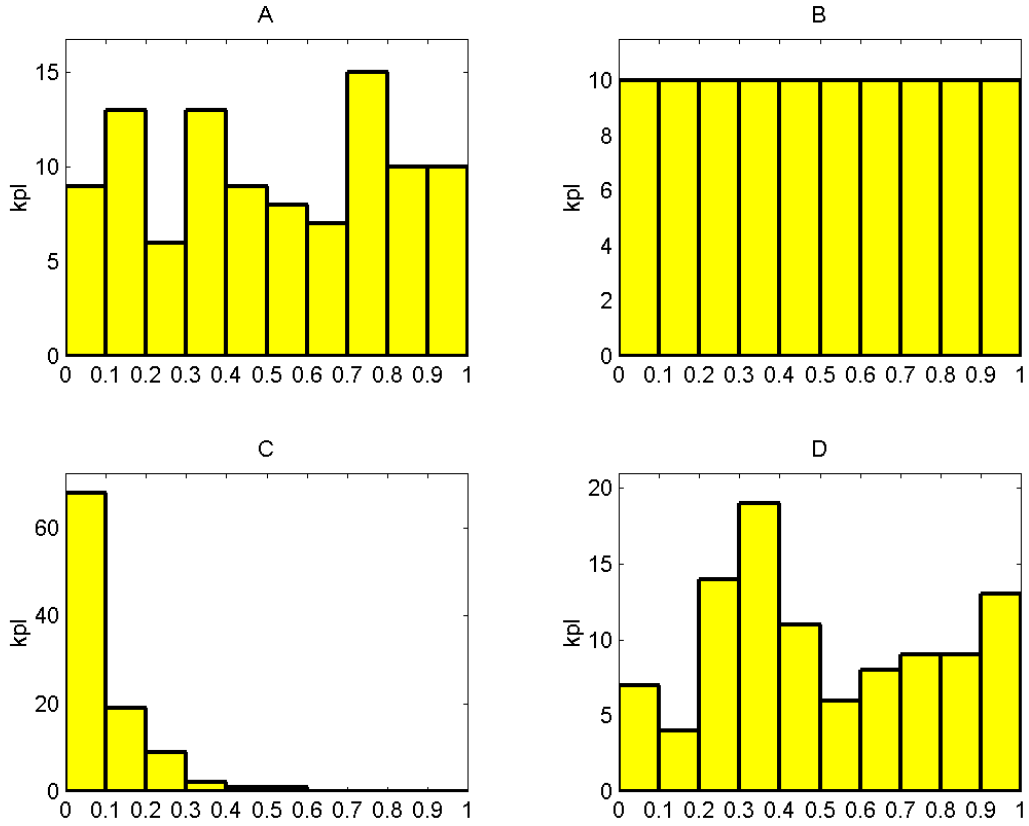


RIIPPUMATON OTOS TASAJAKAUMASTA

Tunnistustehtävä

Kukin näistä histogrammeista on piirretty sadasta luvusta.

Mikä histogrammeista on syntynyt siten, että luvut on arvottu jakaumasta **Tas(0, 1)** ?



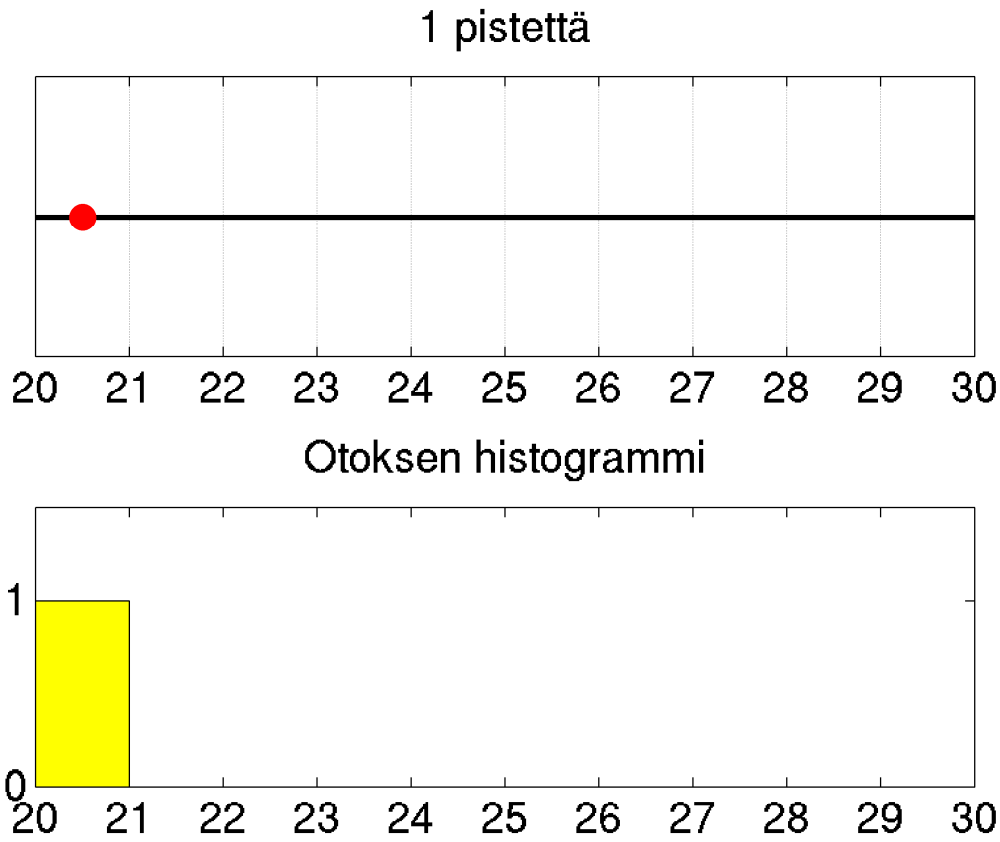
Otos tasajakaumasta

Arvotaan riippumattomia satunnaislukuja
$X_1, X_2, X_3 \ldots \sim$ Tas(20, 30).

Miten ne sijoittuvat?

Piirretään myös 10 pylvään histogrammi, ei **jakaumasta** (se on tasainen), vaan **otoksesta**, ts. mille väleille arvotut luvut osuivat.

(Histogrammi on yhteenveto siitä, missä pisteet *suunnilleen* ovat.)

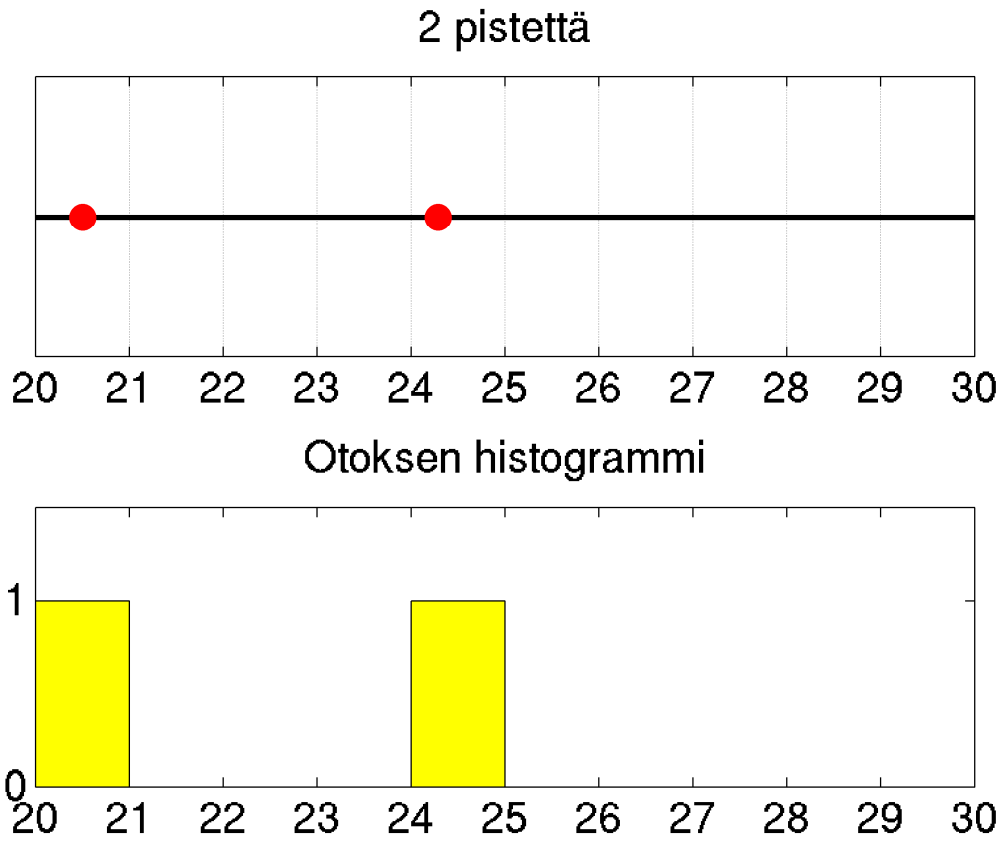


Otos tasajakaumasta

Arvotaan riippumattomia satunnaislukuja
$X_1, X_2, X_3 \ldots \sim$ Tas(20, 30).

Miten ne sijoittuvat?

Piirretään myös 10 pylvään histogrammi, ei **jakaumasta** (joka on tasajakauma), vaan **otoksesta**, ts. mille väleille arvotut luvut osuivat.

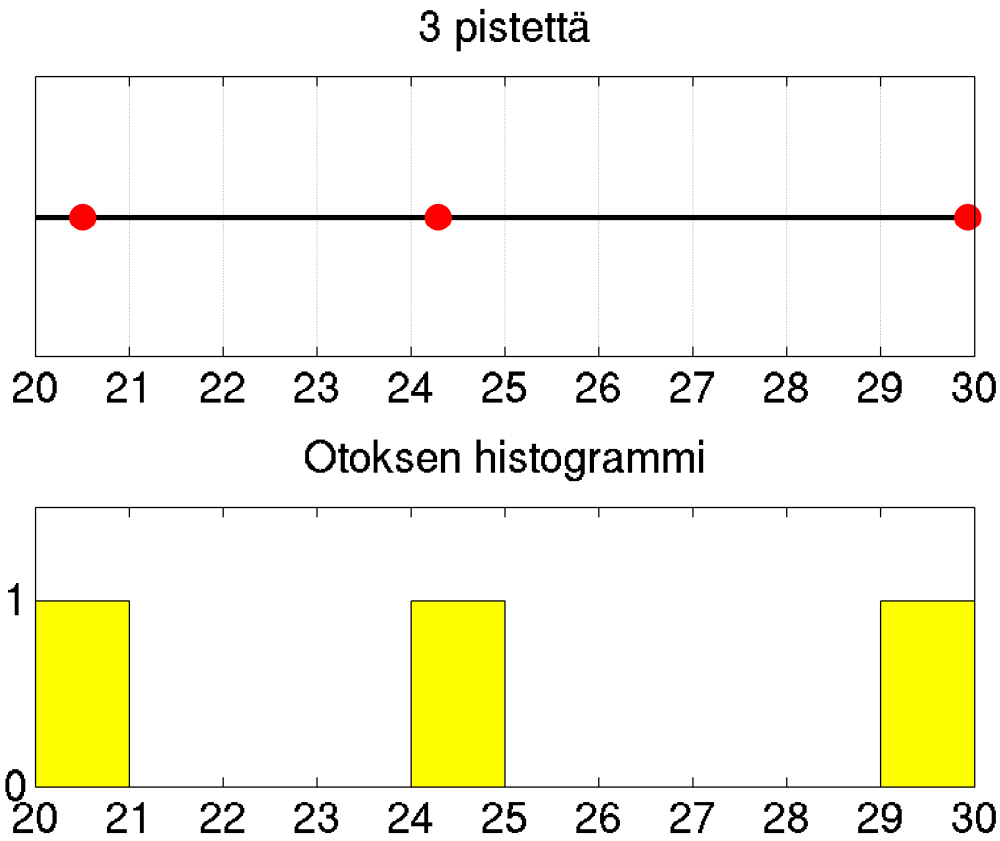


Otos tasajakaumasta

Arvotaan riippumattomia satunnaislukuja
$X_1, X_2, X_3 \ldots \sim$ Tas(20, 30).

Miten ne sijoittuvat?

Piirretään myös 10 pylvään histogrammi, ei **jakaumasta** (joka on tasajakauma), vaan **otoksesta**, ts. mille väleille arvotut luvut osuivat.



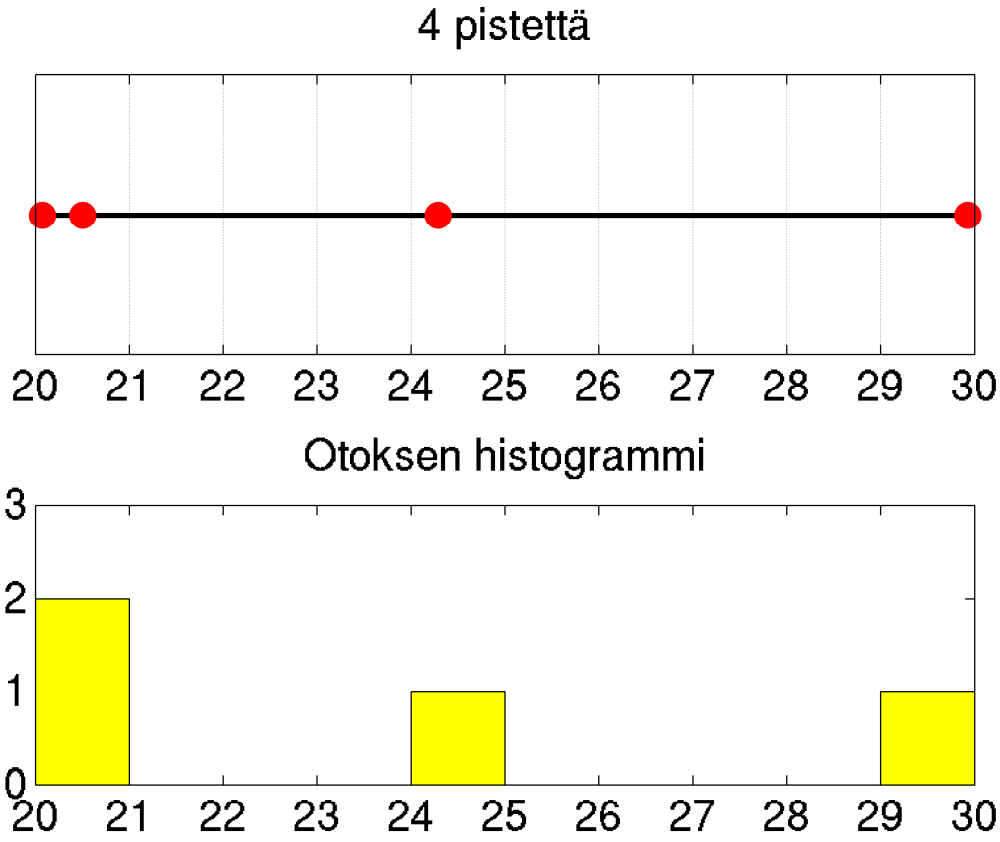
Otos tasajakaumasta

Arvotaan **riippumattomia** satunnaislukuja
$X_1, X_2, X_3 \ldots \sim$ Tas(20, 30).

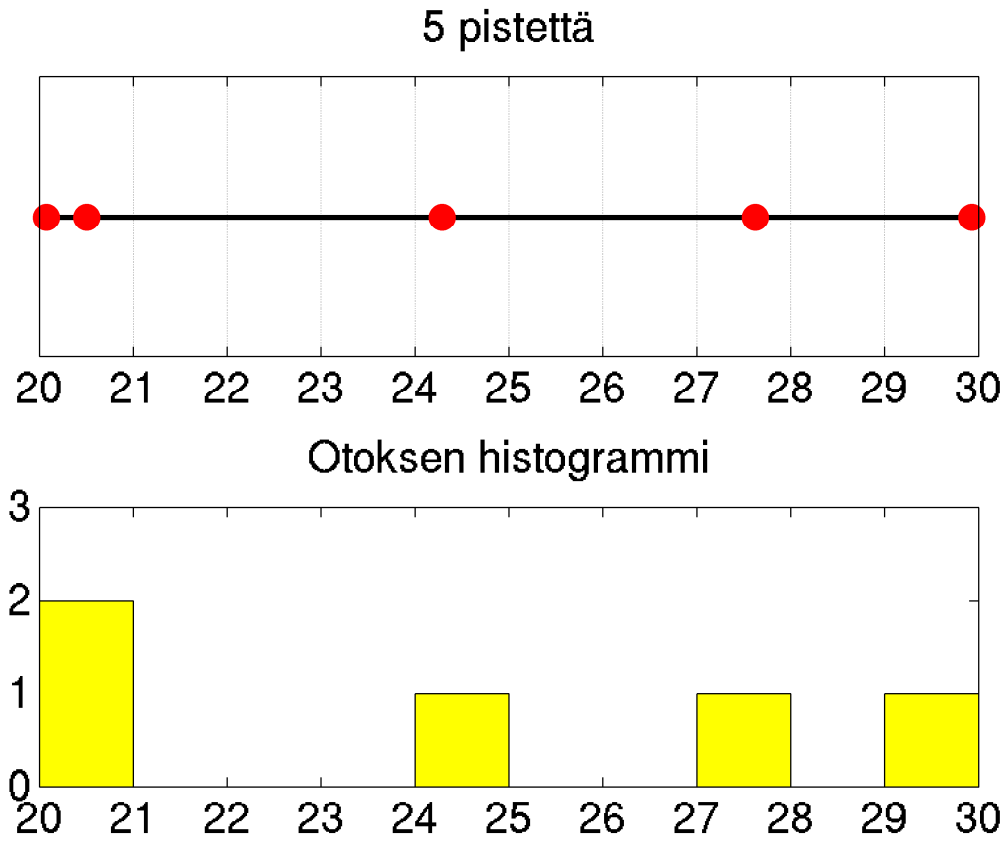
Riippumattomuus:
Aiemmat pisteet eivät vaikuta myöhempiin.

1. piste oli välillä (20,21),
se mitenkään estä 4. pistettä osumasta samalle välille

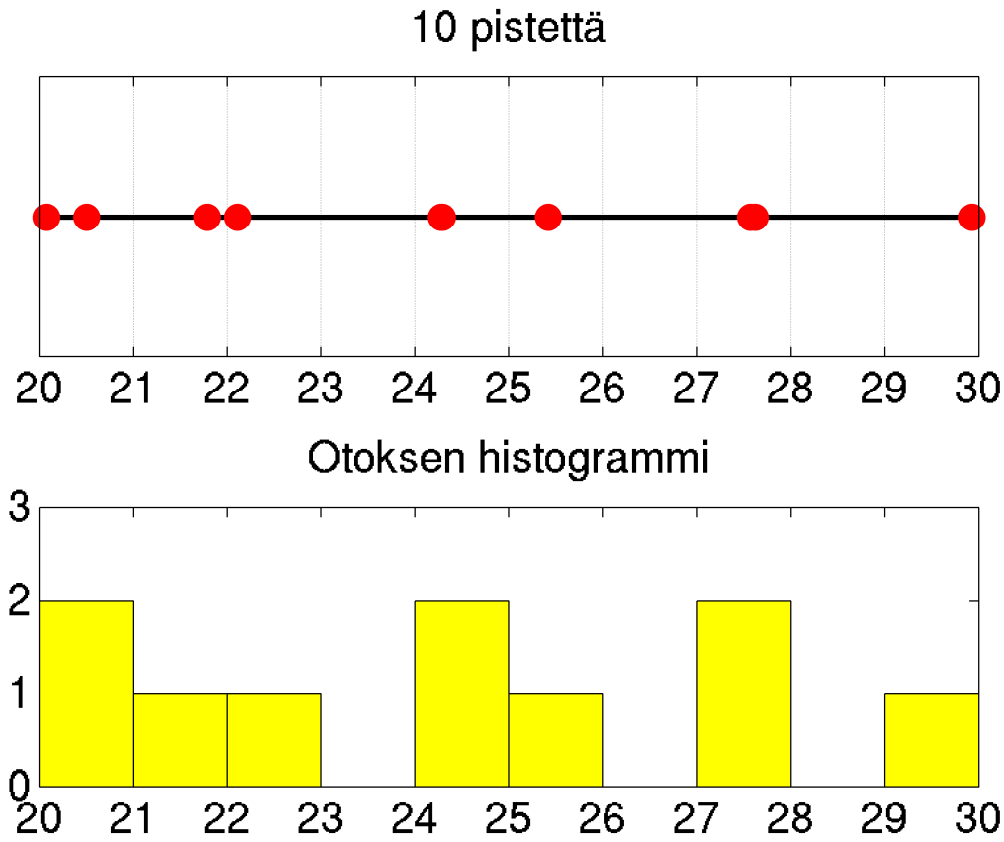
(ei edes lisää eikä vähennä ko. tapahtuman tn:ää, joka on jatkuvasti 1/10)



Otos tasajakaumasta



Otos tasajakaumasta

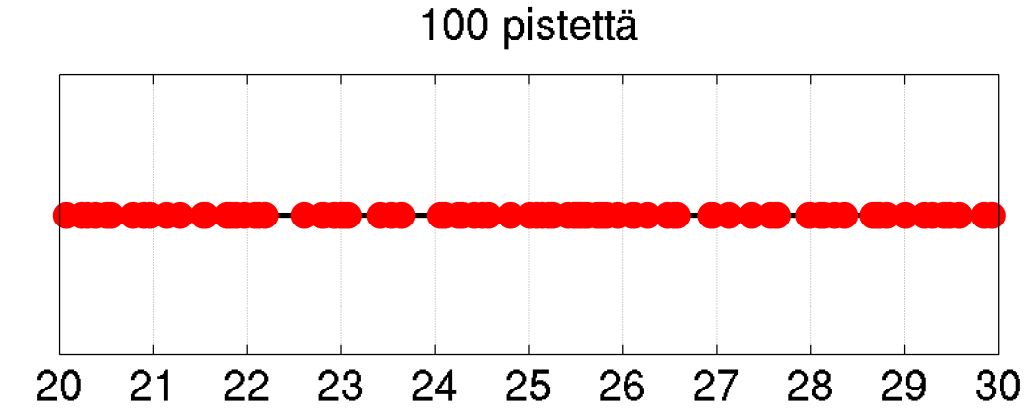


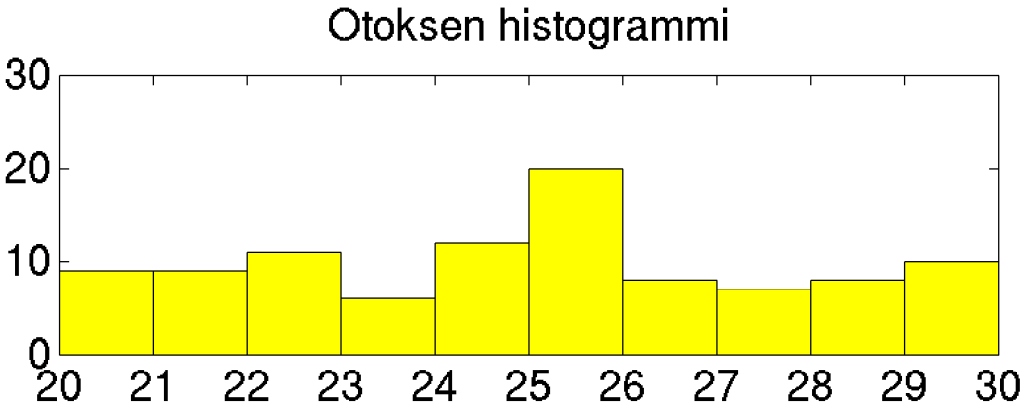
Otos tasajakaumasta

100 pistettä:

Kunkin pylvään
korkeuden odotusarvo
on 10, (miksi?)

mutta toteutuneet korkeudet
vaihtelevat melkoisesti.

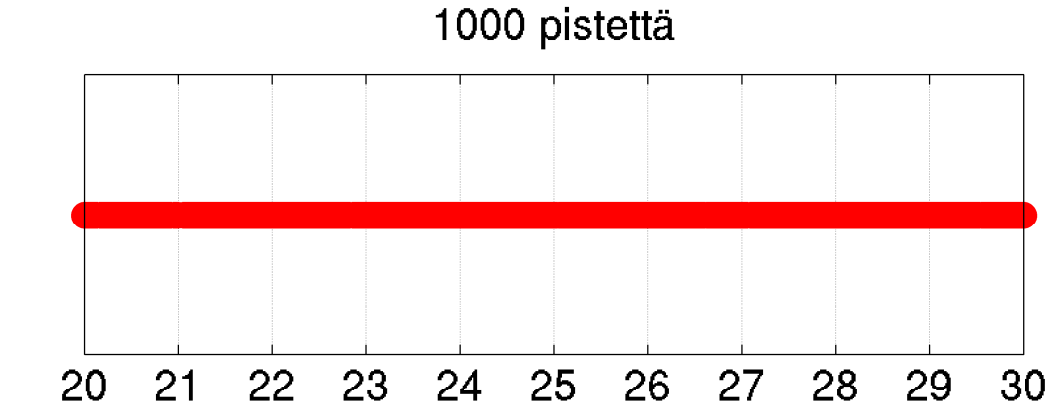


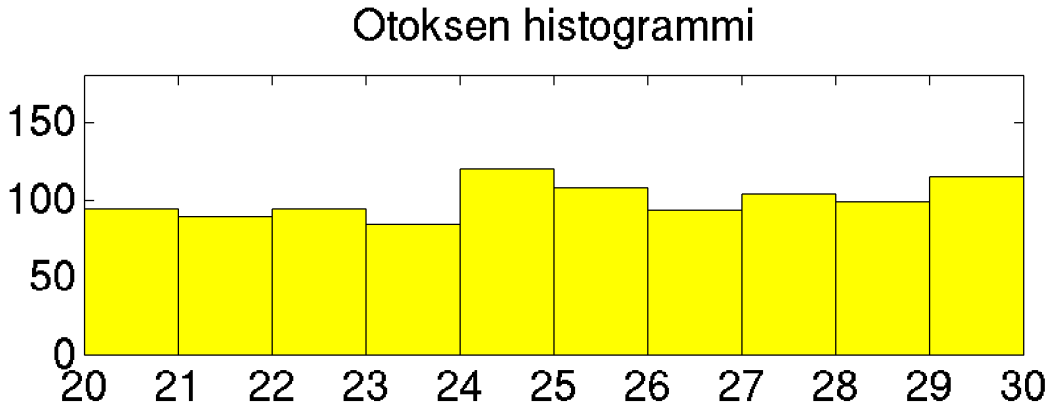


Otos tasajakaumasta

Yksittäisiä pisteitä on jo mahdoton erottaa (jakauma voisi olla joku muu ja näyttäisi samalta)

mutta histogrammista näemme osapuilleen pisteiden jakauman.

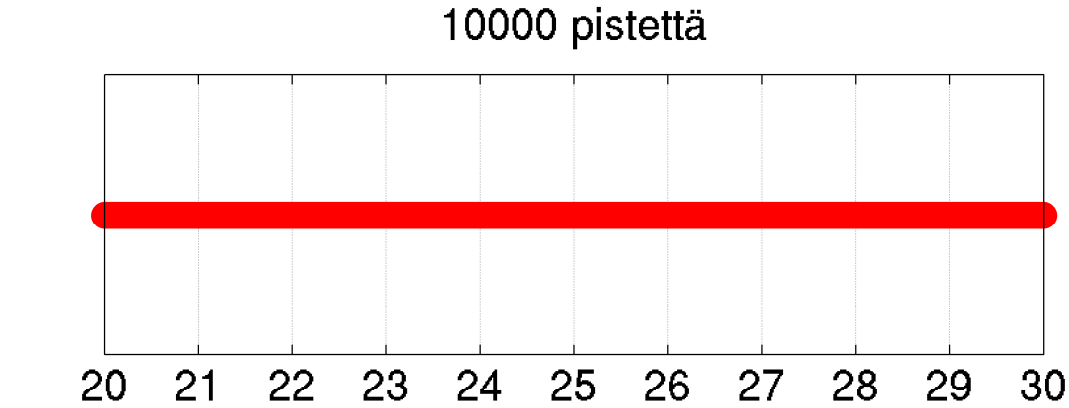


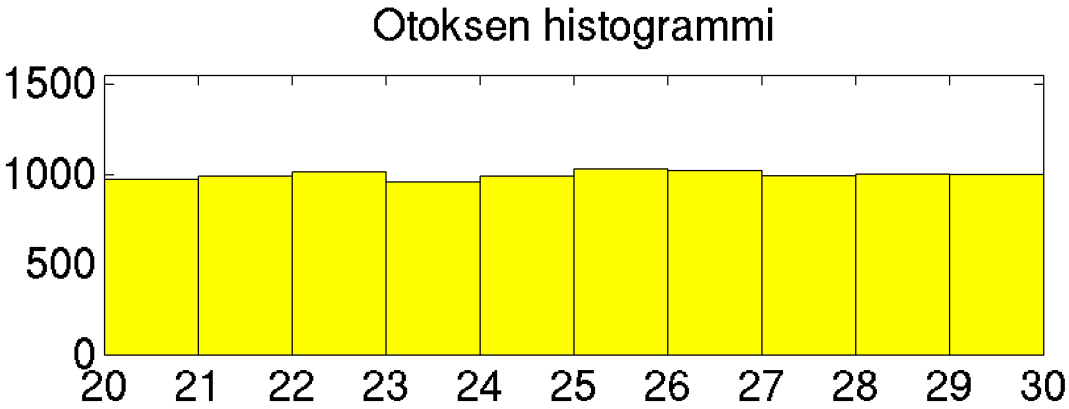


Otos tasajakaumasta

10 000 pistettä:

Melko tasaista.
Otos antaa jo hyvän käsityksen **jakauman** muodosta karkealla tasolla.



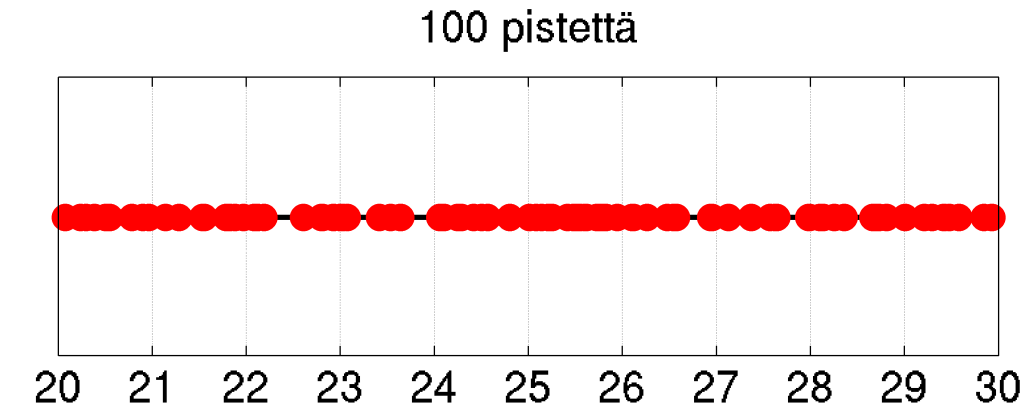


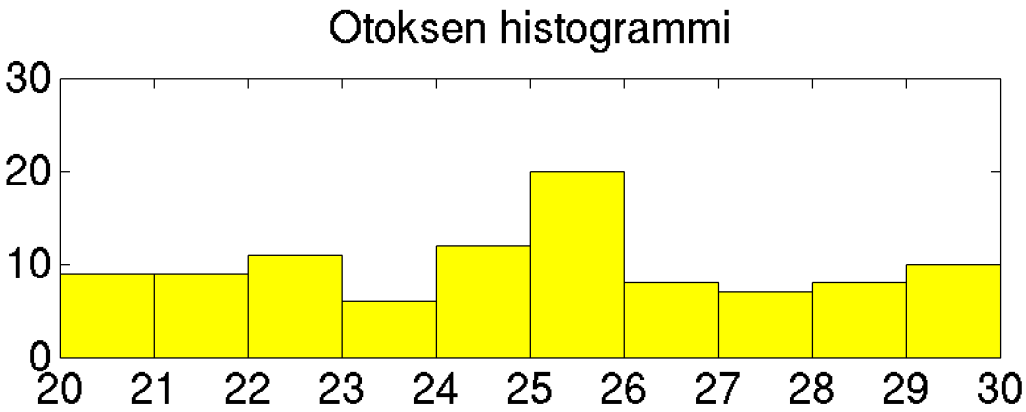
Pylväiden korkeuksien jakauma

Mennään takaisin 100 pisteeseen.
Merk.
Y_i = *i*:nnen pylvään korkeus
= *i*:nnelle jakovälille osuvien pisteiden lukumäärä

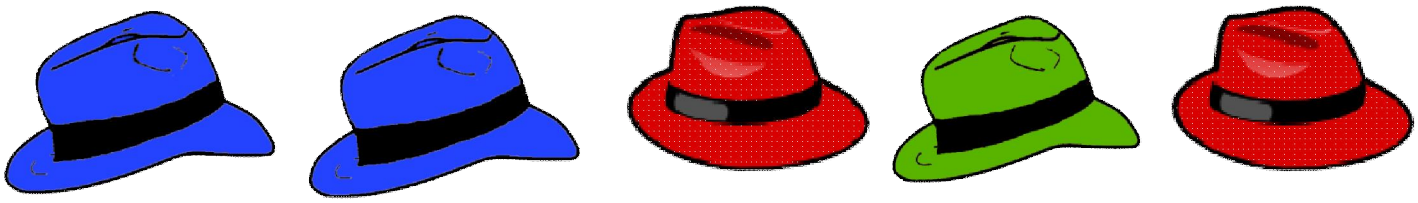
Yksittäisen pylvään korkeus on **binomijakautunut. (miksi?)**

Tn, että histogrammi on täysin tasainen? Huomaa, että pylväiden korkeudet eivät ole riippumattomia. Tarvitaan ns. multinomijakauma.





Ei luultavasti ehditä 7.2. luennolla



MULTINOMIKERROIN JA MULTINOMIKOE

Multinomikerroin

Tuominen 34-36

10 ihmisellä on eriväriset hatut: 3 punaista, 2 sinistä ja 5 vihreää.
Ihmiset asettuvat jonoon.
Monessako eri järjestyksessä värit voivat olla? Emme välitä ihmisistä.

G B E H A J D F I C

Ihmisten jonoja on 10! erilaista. Mutta monessa eri jonossa on sama värijärjestys!

- Jos punaisten (B, E, C) järjestys vaihdetaan keskenään, saadaan sama värijono
- Jos sinisten (J, I) järjestys vaihdetaan keskenään, saadaan sama värijono
- Jos vihreiden (G, H, A, D, F) järjestys vaihdetaan keskenään, saadaan sama värijono

Tuloperiaate: sama värijono saadaan (3!) · (2!) · (5!) eri tavalla

Eri värijonoja on siis

$$\frac{10!}{(3!)\cdot(2!)\cdot(5!)} = \frac{3628800}{6\cdot 2\cdot 120} = 2520 = \binom{10}{3,\,2,\,5}$$

Uusi merkintä: **multinomikerroin**

”Monellako tavalla 10:stä alkiosta voi valita 3 kpl ensimmäiseen osajoukkoon, 2 kpl toiseen ja 5 kpl kolmanteen?” Vrt. binomikerroin.

Menikö oikein?

```
>> P = perms('PPPSSVVVVV');
>> size(P)
ans =
     3628800          10
```

Värit eri ihmisjonoissa
(monessa jonossa sama värijärjestys)

VVVVVSSPPP
VVVVVSSPPP
VVVVVSSPPP
VVVVVSSPPP
VVVVVSSPPP
...
PVVVVVSPSP
PVVVVVSSPP
PVVVVVSSPP

3 628 800 kpl

```
>> U = unique(P, 'rows');
>> size(U)
ans =
        2520          10
```

Erilaiset värijonot

VVVVVPPPSS
VVVVVPPSPS
VVVVVPPSSP
VVVVVPSPPS
VVVVVPSPSP
...
SSPPVVPVVV
SSPPVPVVVV
SSPPPVVVVV

2 520 kpl OK

Multinomikoe

Toistokokeen yleistys: Joka kerralla on **monta** poissulkevaa vaihtoehtoa, joiden tn:t tunnetaan. Kysytään montako kertaa toteutuu mikäkin vaihtoehto, kun kokeita on *n* kpl.

> Suuressa populaatiossa on kolmen puolueen **A**, **B** ja **C** kannattajia osuuksin $p = 0.5$, $q = 0.3$ ja $r = 0.2$.
> Poimitaan populaatiosta umpimähkään $n = 10$ henkilön otos.
>
> Millä todennäköisyydellä saadaan otos, jossa puolueiden kannattajien **lukumäärät** ovat a, b ja c (missä $a + b + c = 10$)?

- Suuri populaatio: approksimoimme otoksen ”takaisinpanolla”, kukin otoksen henkilö on toisista **riippumatta** A:n kannattaja tn:llä p jne.
- Alkeistapauksina mahdolliset 10-jonot puoluekantoja otoksessa – ei symmetriset! Esim. eräiden jonojen todennäköisyyksiä:

P(AAAAAAAAAA)	$= p^{10}$	$\approx 0.000\,977$	
P(AAAABBBBCC)	$= p^4 \cdot q^4 \cdot r^2$	$\approx 0.000\,020$	Miksi pienempi?!

Multinomikoe

Meitä ei kiinnosta, missä järjestyksessä puoluekantoja ilmaantuu otokseen, vain lukumäärät. **Lasketaan yhteen** alkeistapaukset, joissa lukumäärät ovat samat. Tapausten määrä = multinomikerroin!

P(AAAAAAAAAA)	$= p^{10}$	$\approx 0.000\ 977$	Näitä (10×A) on vain yksi
P(AAABABBBCC)	$= p^4 \cdot q^4 \cdot r^2$	$\approx 0.000\ 020$	Näitä alkeistapauksia (4×A, 4×B, 2×C) on $\binom{10}{4,4,2} = 3150$ kpl,
P(BBCAABBAAC)	$= p^4 \cdot q^4 \cdot r^2$	$\approx 0.000\ 020$	
P(AABCCAABBB)	$= p^4 \cdot q^4 \cdot r^2$	$\approx 0.000\ 020$	
...			
P(CCBBBBAAAA)	$= p^4 \cdot q^4 \cdot r^2$	$\approx 0.000\ 020$	tn yht. ≈ **0.064**

...

Multinomikoe

Tässä eri lukumäärien todennäköisyyksiä suuruusjärjestyksessä.

(a,b,c)	**tn**
(5,3,2)	0.085
(6,2,2)	0.071
(6,3,1)	0.071
(4,4,2)	0.064
(5,4,1)	0.064
…	…
(10,0,0)	0.000 977
…	…
(0,0,10)	0.000 000 102
yhteensä	**1**

Tn, että otososuudet ovat **täsmälleen** samat (50%, 30%, 20%) kuin populaatiossa, on vain 0.085

Mutta muutkaan todennäköiset osuudet eivät **paljon** poikkea

Tn saada ”pahasti pielessä” oleva otos on hyvin pieni

Multinomikoe ja multinomijakauma

- *n* riippumatonta koetta, jokaisessa 3 poissulkevaa vaihtoehtoa.
- Joka kerta vaihtoehtojen todennäköisyydet *p*, *q*, *r*.
- Tn, että toteutuneet lukumäärät ovat (*a*, *b*, *c*) on

$$\binom{n}{a,b,c} \cdot p^a \cdot q^b \cdot r^c$$

- Voimme sanoa, että lukumäärät (*a*, *b*, *c*) ovat yhdessä **multinomijakautuneet** parametrein *n* ja (*p*, *q*, *r*).
- Lukumäärät ovat satunnaismuuttujia, ja keskenään **riippuvia**!
  Jos esim. sattuu *a*=*n*, niin on pakko olla *b*=*c*=0. (miksi?)
- Jos vaihtoehtoja on > 3, kaava yleistyy ilmeisellä tavalla.
- Jos vaihtoehtoja on vain 2, kaava palautuu tuttuun binomijakaumaan.