

# MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

## 3B Tilastolliset datajoukot

Emilia Blåsten

Matematiikan ja systeemianalyysin laitos  
Perustieteiden korkeakoulu  
Aalto-yliopisto

Lukuvuosi 2021–2022  
Periodi II

# Sisältö

## Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

Histogrammi

Kahden muuttujan datajoukot

Kvantiilit

Datajoukko vs. yleistäminen

# Mitä tilastotiede on?

- Tilastotiede soveltaa ja kehittää menetelmiä, joita voidaan käyttää tutkittaessa **reaalimaailman** ilmiöitä, joiden tietoihin liittyy **satunnaisuutta tai epävarmuutta**.
  - Epävarmuuden lähteitä on monta: Fysikaalinen satunnaisuus, vajaa tietämys ilmiön lainalaisuuksista, satunnaisotanta, mittausvirheet, puuttuva data . . .
  - Menetelmät perustuvat todennäköisysteorian lainalaisuuksiin.
- Karkea luonnehdinta:
  - T<sub>n</sub>-teoria kertoo, miten jokin prosessi tuottaa dataa.
  - Tilastotiede kertoo, mistä prosessista jokin data on syntynyt.
- Tilastotiedettä voidaan soveltaa aina, kun saatavilla on kvantifioitavaa dataa.
  - Mikä tahansa datajoukko, joka kuvaa jotakin reaalimaailman ilmiötä on potentiaalinen tilastotieteen tutkimuskohde.

# Kaksi näkökulmaa tilastotieteeseen

- Datan **kuvailemisen** (engl. *descriptive statistics*) menetelmiä
  - Taulukot (“raaka data”)
  - Erilaiset kuvat
  - Tunnusluvut, numeeriset “yhteenvedot” (esim. keskiarvo, kvantiilit, korrelaatio)
- Tilastollisen **päätelyn** (engl. *statistical inference*) menetelmiä, joilla pyritään **yleistämään** havaitun datan ulkopuolelle (koko populaatioon tai universaaliin lakiin)
  - Stokastiset mallit
  - Parametrien estimointi
  - Merkitsevyyden testaus

# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

Histogrammi

Kahden muuttujan datajoukot

Kvantiilit

Datajoukko vs. yleistäminen

# Tilastollinen data

Tilastollisen analyysin kohteena oleva data on usein tapana tallettaa taulukkoon eli datakehikkoon (“data frame”), jonka

- rivit vastaavat kohteesta tehtyjä **havaintoja**
- sarakkeet vastaavat tutkittavan ilmiön **muuttujia**

Muuttujat voivat olla laadullisia tai määrällisiä

- **laadullisen** muuttujan arvot jaotellaan luokkiin (esim. 'aurinkoista', 'sateista', 'pilvistä')
- **määrällisen** muuttujan arvot ovat lukuja

## Erilaisia mitta-asteikkoja

- **luokka-asteikko** (nominaaliasteikko): vain joukko eri arvoja  
sukupuoli: {mies, nainen}  
pääaine: {matematiikka, fysiikka, kemia}
- **järjestysasteikko**: luokilla on mielekäs järjestys  
vaatekoko: { XS < S < M < L < XL }  
Likert-asteikko:  
{täysin eri mieltä < eri mieltä < neutraali < samaa mieltä < täysin s. m.  
}
- **numeerinen**: muuttujan arvoilla on aritmeettinen merkitys
  - **intervalliasteikko**: erotuksilla  $x - y$  on merkitystä  
päivämäärät, Celsius-lämpötila
  - **suhdeasteikko**: myös osamäärät  $x/y$  ovat mielekkäitä  
pituus, paino, etäisyys, Kelvin-lämpötila
- Kaikki muuttujat voi *esittää* lukuina, esim. mat=1, fys=2, kem=3, mutta aritmetiikka ei aina ole mielekästä.
- luokka-asteikko=“laatuasteikko” ( $\neq$  “laadullinen tutkimus”)
- tämä ei täysin vastaa erottelua diskreetti/jatkuva. Numeerinen data voi hyvin olla diskreettiä (esim. jonkun lukumäärät)

# Datajoukko

Datajoukko = Järjestetty lista keskenään samantyyppisiä alkioita, esim. lukuja, merkkijonoja tai näistä muodostettuja listoja

Esim. Kurssipalaute: ((12345A, 5, 1, 5), (98759K, 1, 5, 2), (33312K, 4, 4, 3), (23453B, 4, 4, 3), (21453U, 3, 3, 3))

Yksi merkkijonoarvoinen muuttuja (opiskelijanumero) ja kolme lukuarvoista muuttujaa (yleisarvio, työläys, hyödyllisyys)

Opiskelijanumero	Yleisarvio	Työläys	Hyödyllisyys
12345A	5	1	5
98759K	1	5	2
33312K	4	4	3
23453B	4	4	3
21453U	3	3	5

5 havaintoyksikköä, 4 muuttujaa



## Datajoukon keskiarvo ja keskihajonta

Lukuarvoinen yhden muuttujan datajoukko  $\vec{x} = (x_1, \dots, x_n)$

$$\text{Keskiarvo } m(\vec{x}) = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Varianssi } \text{var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - m(\vec{x}))^2$$

$$\text{Keskihajonta } \text{sd}(\vec{x}) = \sqrt{\text{var}(\vec{x})}$$

Esim.  $\vec{y} = (0, 0, 1, 1, 2, 2)$

$$m(\vec{y}) = \frac{1}{6} (0 + 0 + 1 + 1 + 2 + 2) = 1$$

$$\text{var}(\vec{y}) = \frac{1}{6} \left( (0-1)^2 + (0-1)^2 + (1-1)^2 + (1-1)^2 + (2-1)^2 + (2-1)^2 \right) = \frac{2}{3}$$

$$\text{sd}(\vec{y}) = \sqrt{\frac{2}{3}} \approx 0.8165$$

**Huom:** Joskus varianssin laskennassa käytetään jakajaa  $n - 1$  eikä  $n$ . Tämä liittyy tilanteeseen, jossa datan varianssilla halutaan estimoida suuremman populaation varianssia. Tästä lisää myöhemmin.

## Esimerkki

Laske keskiarvo ja keskihajonta seuraaville datajoukoille:

$$\vec{x} = (1, 1, 1, 1, 1),$$

$$\vec{y} = (0, 0, 1, 1, 2, 2),$$

$$\vec{z} = (0, 2, 0, 2, 0, 2, 0, 2, 0, 2),$$

$$\vec{w} = (\underbrace{0, 0, 0, 0, \dots, 0, 0, 0, 0}_{666666 \text{ kpl}}, \underbrace{1000000, 0, 0, \dots, 0, 0}_{333333 \text{ kpl}}).$$

Datajoukko	Keskiarvo	Keskihajonta
$\vec{x}$	1	0.0000
$\vec{y}$	1	0.8165
$\vec{z}$	1	1.0000
$\vec{w}$	1	999.9995

Keskiarvo ja keskihajonta ovat datan *yhteenvetoja*, ja kertovat datasta vain jotakin, eivät kaikkea. (Kuten todennäköisyysjakaumissakin.)

# Tunnuslukujen laskeminen

Merkintä	Nimitys	R	Python	Excel
$m(\bar{x})$	Keskiarvo	mean()	np.mean()	AVERAGE()
$sd(\bar{x})$	Keskihajonta	sqrt(1-1/n)*sd()	np.std()	STDEV.P()
$sd_s(\bar{x})$	Otoskeskihajonta	sd()	np.std(,ddof=1)	STDEV.S()
$var(\bar{x})$	Varianssi	(1-1/n)*var()	np.var()	VAR.P()
$var_s(\bar{x})$	Otosvarianssi	var()	np.var(,ddof=1)	VAR.S()
$cov(\bar{x}, \bar{y})$	Kovarianssi	(1-1/n)*cov()	np.cov(,ddof=0)[0][1]	COVARIANCE.P()
$cov_s(\bar{x}, \bar{y})$	Otoskovarianssi	cov()	np.cov(,ddof=1)[0][1]	COVARIANCE.S()
$cor(\bar{x}, \bar{y})$	Korrelaatio	cor()	np.corrcoef()[0][1]	CORREL()
$q_{0.5}(\bar{x})$	Mediaani	median()	np.median()	MEDIAN()
$q_{0.25}(\bar{x})$	Alakvartiili	quantile(, .25)	np.quantile(, .25)	PERCENTILE.INC(, .25)
$q_{0.75}(\bar{x})$	Yläkvartiili	quantile(, .75)	np.quantile(, .75)	PERCENTILE.INC(, .75)

$$\text{Otoskeskihajonta } sd_s(\bar{x}) = \left(1 - \frac{1}{n}\right)^{1/2} sd(\bar{x})$$

**Huom.** Joissakin ohjelmissa voit itse valita, lasketaanko ns. varianssi vai otosvarianssi eli onko jakajana  $n$  vai  $n - 1$ . Vaihtoehtoisesti luvun voi aina skaalata.

# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

**Empiirinen jakauma**

Histogrammi

Kahden muuttujan datajoukot

Kvantiilit

Datajoukko vs. yleistäminen

# Esiintyvyytaulukko

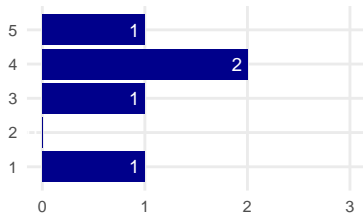
Arvon  $x$  esiintyvyys eli frekvenssi

$$n_{\vec{x}}(x) = \#\{i : x_i = x\}$$

on datajoukossa  $\vec{x} = (x_1, \dots, x_n)$  arvoltaan  $x$  olevien alkoiden lukumäärä

Kurssipalautteen muuttujaa "Yleisarvio" vastaavan datajoukon (5, 1, 4, 4, 3) esiintyvyytaulukko ja palkkikaavio:

$x$	1	2	3	4	5
$n_{\vec{x}}(x)$	1	0	1	2	1



# Suhteelliset esiintyvyydet

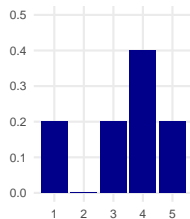
Arvon  $x$  **suhteellinen esiintyvyys**

$$f_{\vec{x}}(x) = \frac{n_{\vec{x}}(x)}{n} = \frac{\#\{j : x_j = x\}}{n}$$

on datajoukossa arvoltaan  $x$  olevien alkioden suhteellinen osuus

Kurssipalautteen muuttujaa "Yleisarvio" vastaavan datajoukon (5, 1, 4, 4, 3) esiintyvyydestaulukko ja **pylväskaavio** (engl. *bar chart*):

$x$	1	2	3	4	5
$f_{\vec{x}}(x)$	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$



Huomataan:  $\sum_x f_{\vec{x}}(x) = 1 \implies f_{\vec{x}}(x)$  on eräs todennäköisyysjakauma!  
 $f_{\vec{x}}(x)$  on datajoukon  $\vec{x}$  **empiirinen jakauma**.

# Empiirinen jakauma

## Lause

Datajoukosta  $\vec{x} = (x_1, \dots, x_n)$  tasaisen satunnaisesti valittu alkio  $X$  on diskreetti satunnaismuuttuja, joka noudattaa datajoukon  $\vec{x}$  empiiristä jakaumaa tiheysfunktiona  $f_X(x) = f_{\vec{x}}(x)$  ja toteuttaa

$$\mathbb{E}(X) = m(\vec{x}), \quad (1)$$

$$\text{SD}(X) = \text{sd}(\vec{x}), \quad (2)$$

$$\text{Var}(X) = \text{var}(\vec{x}). \quad (3)$$

Lisäksi mielivaltaiselle funktiolle  $g$  pätee

$$\mathbb{E}[g(X)] = \frac{1}{n} \sum_{i=1}^n g(x_i). \quad (4)$$

## Esimerkki

Määritä empiirinen jakauma ja laske sen avulla keskiarvo ja keskihajonta datajoukolla  $\vec{y} = (0, 0, 1, 1, 2, 2)$ .

Arvojen suhteelliset esiintyvyydet ovat

$y$	0	1	2
$f_{\vec{y}}(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Tiheysfunktion  $f_{\vec{y}}(y)$  mukaan jakautuneelle satunnaismuuttujalle  $Y$  pätee

$$\mathbb{E}(Y) = \sum_{y=0}^2 y f_{\vec{y}}(y) = 0 \times \frac{1}{3} + 1 \times \frac{1}{3} + 2 \times \frac{1}{3} = 1,$$

$$\text{Var}(Y) = \sum_{y=0}^2 (y-1)^2 f_{\vec{y}}(y) = (0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3} = \frac{2}{3}$$

$$\implies m(\vec{y}) = \mathbb{E}(Y) = 1$$

$$\implies \text{sd}(\vec{y}) = \sqrt{\text{var}(\vec{y})} = \sqrt{\text{Var}(Y)} = \sqrt{\frac{2}{3}} \approx 0.8165$$



# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

**Histogrammi**

Kahden muuttujan datajoukot

Kvantiilit

Datajoukko vs. yleistäminen

# Luokittelu ja histogrammi

Esim. Suomalaisten ikärakenne 31.12.2015.

$n = 5\,487\,308$  miljoonaa datapistettä

Ei ole järkeä piirtää jokaista pistettä kuvaajaan

Jaetaan datapisteet luokkiin.

Ikä (v)	Lukumäärä
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

# Luokittelu ja histogrammi

Histogrammi piirretään yleensä näin:

- Yksi palkki per luokka
- Palkin leveys = luokkavälin leveys (yksikkönä vuosi)
- Palkin korkeus = datapisteiden suhteellinen osuus jaettuna palkin leveydellä (yksikkönä % per vuosi)

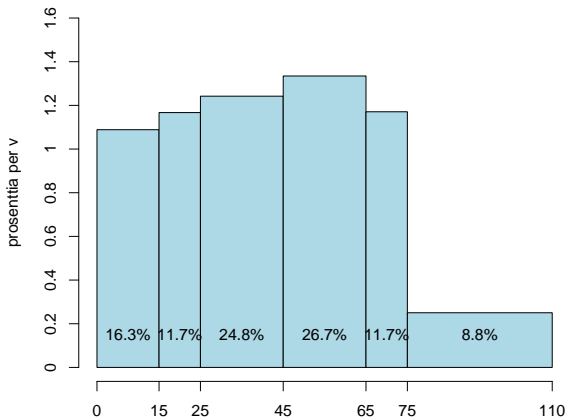
Esim: Suomalaiset

- 1. palkki käsittää suomalaiset, joiden ikä on 0–14 vuotta
- 1. palkin leveys = 15 v
- Datapisteiden lkm luokassa 1 on 896023 ja suhteellinen osuus  $896023/5487308 \approx 16.3\%$
- Palkin korkeus =  $16.3/15 \approx 1.09$  (yksikkönä % per vuosi).

Huom: Luokkavälit voivat olla samanleveyisiä, mutta niiden ei tarvitse olla.

# Luokittelu ja histogrammi

Suomen väestörakenne ikäluokittain 31.12.2015 [Lähde: Tilastokeskus]



Ikä (v)	Lukumäärä
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

Palkin leveys  $\propto$  luokkavälin leveys. Entä jos palkit olisivat samanlevyisiä?  
Arvioi mikä osuus väestöstä kuuluu ikäluokkaan 13–14 v.  
Entä ikäluokkaan 109–110 v? Kuinka luotettavia arviot ovat?

# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

Histogrammi

**Kahden muuttujan datajoukot**

Kvantiilit

Datajoukko vs. yleistäminen

## Kahden muuttujan datajoukko

Kahden muuttujan datajoukko = järjestetty lista pareja

$$\vec{x}\vec{y} = ((x_1, y_1), \dots, (x_n, y_n)).$$

Voidaan tulkita myös parina  $(\vec{x}, \vec{y})$ , jossa  $\vec{x} = (x_1, \dots, x_n)$  ja  $\vec{y} = (y_1, \dots, y_n)$  ovat samankokoisia yhden muuttujan datajoukkoja

Kurssipalautteen muuttujat "Yleisarvio" ja "Hyödyllisyys" voidaan koostaa datajoukoksi  $((5,5), (1,2), (4,3), (4,3), (3,3))$

Tunnuslukuja:  $m(\vec{x}), m(\vec{y}), sd(\vec{x}), sd(\vec{y})$

Nämä eivät kerro mitään muuttujien riippuvuuksista

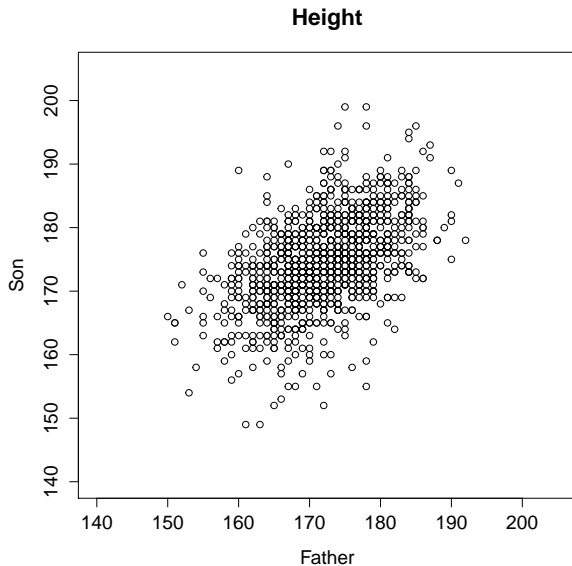
Muuttujien yhteisvaihtelua kuvaavat kovarianssi ja korrelaatio.

$$\text{cov}(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - m(\vec{x}))(y_i - m(\vec{y}))$$

$$\text{cor}(\vec{x}, \vec{y}) = \frac{\text{cov}(\vec{x}, \vec{y})}{sd(\vec{x}) sd(\vec{y})}$$

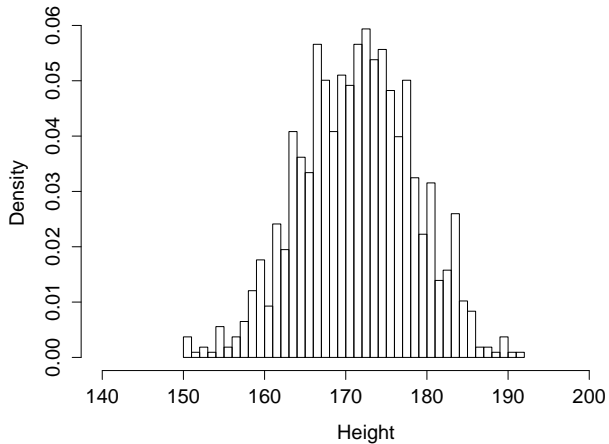


# Hajontakuvio (scatterplot)

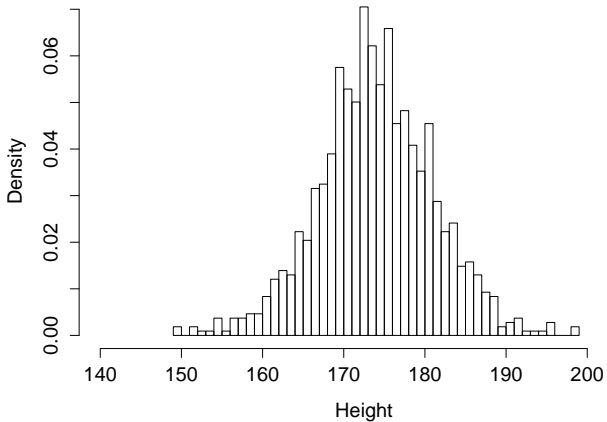




### Histogram of Fathers



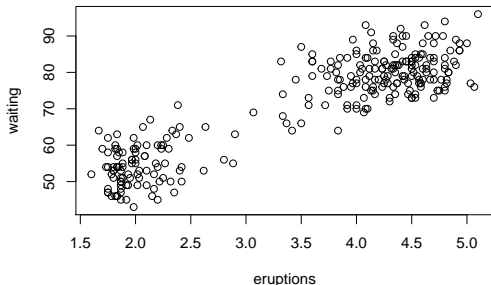
## Histogram of Sons



## Esimerkki: Old Faithful -geysirin purkaukset

Hajontakuvio, 272 purkausta, *Old Faithful* (Yellowstone).

2 muuttujaa: Purkauksen kesto ja väliaika seuraavaan purkaukseen.



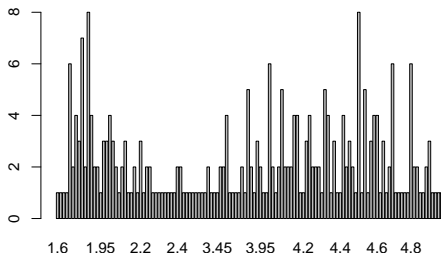
Datan silmäilystä (*eyeballing*) on hyvä aloittaa, jo se voi paljastaa ilmiön olennaisia piirteitä.

Kokeile R:ssä `faithful` ja `help("faithful")`

## Old Faithful: yhden muuttujan pylväskaavio...

Yritetään laskea montako kertaa kukin *eri arvo* esiintyy (purkauksen pituudessa) ja piirretään lukumääristä pylväskaavio

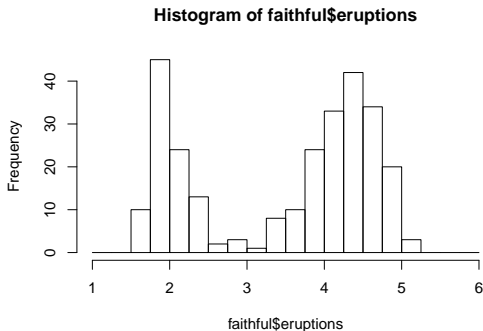
$x$	1.6	1.667	1.7	1.733	1.75	...	5.1
$n_{\bar{x}}(x)$	1	1	1	1	6	...	1



Ei kovin informatiivista.

## Old Faithful: histogrammi

**Ryhmitellään** havainnot 0.25 minuutin väleille kuten [2.00, 2.25).  
Piirretään *väleille* osumisen lukumäärät.



Kokeile itse erilaisia jakovälejä. Mitä tapahtuu hyvin pienillä tai isoilla jakoväleillä?

# Esiintyvyyksien ristitaulukko

Arvoparin  $(x, y)$  esiintyvyys (engl. frequency)

$$n_{x\bar{y}}(x, y) = \#\{i : x_i = x \text{ ja } y_i = y\}$$

on datajoukossa arvoltaan  $(x, y)$  olevien alkioden lukumäärä.

Kurssipalautteen muuttujat "Yleisarvio" ja "Hyödyllisyys" voidaan koostaa datajoukoksi  $((5,5), (1,2), (4,3), (4,3), (3,3))$

	y					
x	1	2	3	4	5	Yht
1	0	1	0	0	0	1
2	0	0	0	0	0	0
3	0	0	1	0	0	1
4	0	0	2	0	0	2
5	0	0	0	0	1	1
Yht	0	1	3	0	1	

# Suhteellisten esiintyvyyksien ristitaulukko

Arvo  $\vec{xy}$   $(x, y)$  suhteellinen esiintyvyys

$$f_{\vec{xy}}(x, y) = \frac{\#\{i : x_i = x \text{ ja } y_i = y\}}{n}$$

on datajoukossa arvoltaan  $(x, y)$  olevien alkioden suhteellinen osuus

	<u>y</u>					
<u>x</u>	1	2	3	4	5	Yht
1	0	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$
2	0	0	0	0	0	0
3	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$
4	0	0	$\frac{2}{5}$	0	0	$\frac{2}{5}$
5	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$
Yht	0	$\frac{1}{5}$	$\frac{3}{5}$	0	$\frac{1}{5}$	

$\sum_{x,y} f_{\vec{xy}}(x, y) = 1 \implies f_{\vec{xy}}(x, y)$  on eräs todennäköisyysjakauma.  
 $f_{\vec{xy}}(x, y)$  on datajoukon  $\vec{xy}$  **empiirinen yhteisjakauma**.

# Empiirinen yhteisjakauma

## Lause

Datajoukosta  $\vec{xy} = ((x_1, y_1), \dots, (x_n, y_n))$  *tasaisen satunnaisesti valittu pari*  $(X, Y)$  on diskreetti satunnaismuuttuja, joka noudattaa datajoukon  $\vec{xy}$  empiiristä jakaumaa tiheysfunktiona  $f_{X,Y}(x, y) = f_{\vec{xy}}(x, y)$  ja toteuttaa

$$\begin{aligned}\mathbb{E}(X) &= m(\vec{x}), & \mathbb{E}(Y) &= m(\vec{y}), \\ \text{SD}(X) &= \text{sd}(\vec{x}), & \text{SD}(Y) &= \text{sd}(\vec{y}), \\ \text{Var}(X) &= \text{var}(\vec{x}), & \text{Var}(Y) &= \text{var}(\vec{y}),\end{aligned}\tag{5}$$

sekä

$$\text{Cor}(X, Y) = \text{cor}(\vec{x}, \vec{y}),\tag{6}$$

$$\text{Cov}(X, Y) = \text{cov}(\vec{x}, \vec{y}).\tag{7}$$

Lisäksi mielivaltaiselle kahden muuttujan funktiolle  $g$  pätee

$$\mathbb{E}[g(X, Y)] = \frac{1}{n} \sum_{i=1}^n g(x_i, y_i).\tag{8}$$



# Tunnuslukujen laskeminen

Merkintä	Nimitys	R	Python	Excel
$m(\vec{x})$	Keskiarvo	<code>mean()</code>	<code>np.mean()</code>	<code>AVERAGE()</code>
$sd(\vec{x})$	Keskihajonta	<code>sqrt(1-1/n)*sd()</code>	<code>np.std()</code>	<code>STDEV.P()</code>
$sd_s(\vec{x})$	Otoskeskihajonta	<code>sd()</code>	<code>np.std( ,ddof=1)</code>	<code>STDEV.S()</code>
$var(\vec{x})$	Varianssi	<code>(1-1/n)*var()</code>	<code>np.var()</code>	<code>VAR.P()</code>
$var_s(\vec{x})$	Otosvarianssi	<code>var()</code>	<code>np.var( ,ddof=1)</code>	<code>VAR.S()</code>
$cov(\vec{x}, \vec{y})$	Kovarianssi	<code>(1-1/n)*cov()</code>	<code>np.cov( ,ddof=0)[0][1]</code>	<code>COVARIANCE.P()</code>
$cov_s(\vec{x}, \vec{y})$	Otoskovarianssi	<code>cov()</code>	<code>np.cov( ,ddof=1)[0][1]</code>	<code>COVARIANCE.S()</code>
$cor(\vec{x}, \vec{y})$	Korrelaatio	<code>cor()</code>	<code>np.corrcoef()[0][1]</code>	<code>CORREL()</code>
$q_{0.5}(\vec{x})$	Mediaani	<code>median()</code>	<code>np.median()</code>	<code>MEDIAN()</code>
$q_{0.25}(\vec{x})$	Alakvartiili	<code>quantile( ,.25)</code>	<code>np.quantile( ,.25)</code>	<code>PERCENTILE.INC( ,.25)</code>
$q_{0.75}(\vec{x})$	Yläkvartiili	<code>quantile( ,.75)</code>	<code>np.quantile( ,.75)</code>	<code>PERCENTILE.INC( ,.75)</code>

$$\text{Otoskeskihajonta } sd_s(\vec{x}) = \left(1 - \frac{1}{n}\right)^{1/2} sd(\vec{x})$$

# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

Histogrammi

Kahden muuttujan datajoukot

**Kvantiilit**

Datajoukko vs. yleistäminen

# Järjestystunnuslukuja

Järjestetyn muuttujan (määrällinen tai järjestetty laadullinen) havainnoista  $\vec{x} = (x_1, \dots, x_n)$ , voidaan laskea tason  $p \in (0, 1)$  kvantiili  $Q(p)$ , eli piste, jonka alapuolella on (suunnilleen) osuus  $p$  havainnoista.

Esim.

- $Q(0.25)$  on **alakvantiili**, sen alla 25% havainnoista
- $Q(0.5)$  on **mediaani**, sen alla 50% havainnoista
- $Q(0.75)$  on **yläkvantiili**, sen alla 75% havainnoitsa

R: `quantile(x,p)`, `summary(x)`, `median(x)`

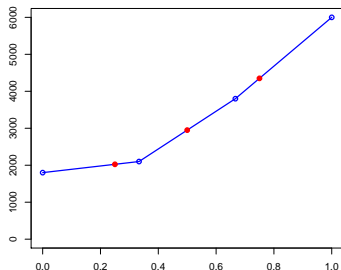
Äärellisellä datalla vain “suunnilleen”, koska esim. 9:stä havainnosta ei voi ottaa tasan puolta. Tähän on erilaisia ratkaisuja, esim. seuraavaksi esitettävä kvantiilifunktio.

# Kvantiilifunktio

Datajoukon  $(x_1, \dots, x_n)$  kvantiilifunktio voidaan määrittää näin:

- Järjestetään datapisteet muotoon  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
- Jaetaan vaaka-akselin yksikköväli tasamittaisiin väleihin, reunapisteinä luvut  $p_k = (k - 1)/(n - 1)$ ,  $k = 1, \dots, n$
- Piirretään tasoon pisteet  $(p_k, x_{(k)})$  ja yhdistetään ne viivoilla

Esim. Neljä bruttopalkkaa (eur/kk): 3800, 1800, 2100, 6000



Kvartiilit = Kvantiilifunktion arvot pisteissä 0.25, 0.50, 0.75

## Esimerkki: Kolme datajoukkoa

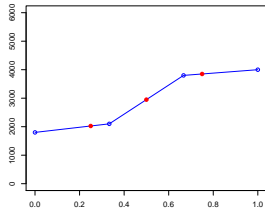
Esim. Piirrä seuraavien datajoukkojen kvantiilifunktiot ja määritä niiden mediaanit ja keskiarvot:

$$\vec{x} = (1800, 2100, 3800, 4000)$$

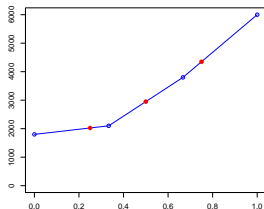
$$\vec{y} = (1800, 2100, 3800, 6000)$$

$$\vec{z} = (1800, 2100, 3800, 6000, 6000)$$

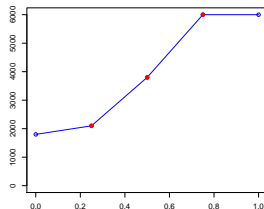
(Huom. Viimeisen joukon koko ei ole neljällä jaollinen.)



$$Q_x(0.50) = 2950,$$
$$m(x) = 2925$$



$$Q_y(0.50) = 2950,$$
$$m(x) = 3425$$



$$Q_z(0.50) = 3800,$$
$$m(x) = 3940$$

# Sisältö

Johdanto

Deskriptiivistä tilastotiedettä

Empiirinen jakauma

Histogrammi

Kahden muuttujan datajoukot

Kvantiilit

Datajoukko vs. yleistäminen

# Datajoukko vs. yleistäminen

Havaitun datajoukon tarkoituksena on usein *esittää* yleisempää ilmiötä, “populaatiota”.

data	populaatio
Pearsonin 1000 isää ja poikaa	Kaikki isät ja pojat (missä? milloin?)
1000 gallup-vastausta	5 miljoonan suomalaisen mielipide (nyt)
272 geysirin purkausta	Kaikki Old F:n purkaukset (menneet/tulevat?)
Lääkkeen vaikutus 30 potilaalla	Vaikutus tulevilla potilailla
100 nopanheitto tulosta	Potentiaalinen ääretön heittojono

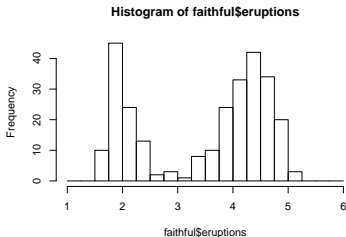
**Populaatio** on tilastotieteen terminologiaa, ja tarkoittaa

- mistä/miten data on syntynyt  
(**generoiva mekanismi**; **datalähde**)
- se mitä datan perusteella yritetään ymmärtää

“Populaatio” ei välttämättä ole mikään konkreettinen kokoelma (esim. ihmisiä).

# Old Faithful vielä kerran

Olemme *havainneet* of 272 eruption lengths. Fysikaalinen mekanismi on ehkä monimutkianen, mutta ajatellaan pituuksien käyttäytyvän **kuten** satunnaismuuttuja, jolla on eräs **jakauma**  $f$ . Mutta mikä jakauma?



Voidaan ajatella, että “kaikki” (historian aikana *toteutuvat* tai fyysikaalisesti *mahdolliset*) purkauspituudet muodostavat “populaation” tai generoivan jakauman, josta havaitut pituudet ovat satunnainen otos.

Empiirinen jakauma **approksimoi** generoivaa jakaumaa. **Miksi?**

Vastaus: Ajattele esim. tapahtuman  $\{2.0 \leq X < 2.25\}$  todennäköisyyttä ja suurten lukujen lakia. Tiedämme, että *suhteellinen esiintyvyys*  $\approx$  *todennäköisyys*.



Seuraavalla kerralla puhutaan parametrien estimoinnista . . .