

MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

3B Tilastolliset datajoukot

Jukka Kohonen

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2022–2023
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 tms. tahallinen satunnaisuus, mittausvirheet, puuttuva data . . .
 - Menetelmät perustuvat todennäköisysteorian lainalaisuuksiin.
- Karkea luonnehdinta:
 - T_n-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:
{täys. eri mieltä < eri mieltä < neutraali < samaa mieltä < täys. 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)), ((223344, 5, 5, 5)), ((98313A, 5, 5, 5))

Opiskelijanumero	Yleisarvio	Työläys	Hyödyllisyys
12345A	5	1	5
98759K	1	5	2
33312K	4	4	3
23453B	4	4	3
21453U	3	3	5
223344	5	5	5
98313A	5	5	5

7 havaintoyksikköä, 4 muuttujaa (merkkijono ja kolme lukua)

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:

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

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

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

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

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 *yhteenvedoja*, 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	$\sqrt{(1-1/n)*sd()^2}$	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

Esiintyvyystaulukko

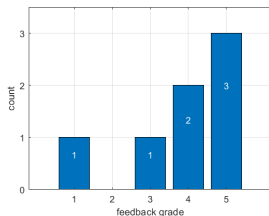
Arvon x esiintyvyys eli frekvenssi

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

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

Kurssipalautteen muuttujaa “Yleisarvio” vastaavan datajoukon $(5, 1, 4, 4, 3, 5, 5)$ esiintyvyystaulukko ja pylväskaavio:

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



Suhteelliset esiintyvyydet

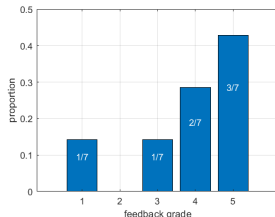
Arvon x **suhteellinen esiintyvyys**

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

on datajoukossa arvoltaan x olevien alkioden suhteellinen osuus

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

x	1	2	3	4	5
$f_{\vec{x}}(x)$	$\frac{1}{7}$	0	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{3}{7}$



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

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

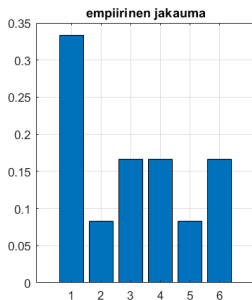
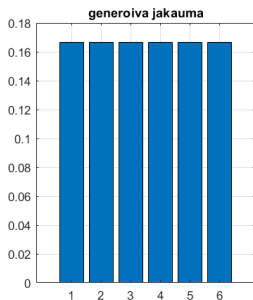
$$\text{Var}(Y) = \sum_{y=0}^2 (y-1)^2 f_{\vec{y}}(y) = (0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3} = \frac{2}{3}$$

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

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

Generoiva ja empiirinen jakauma

Heitetään reilua noppaa 12 kertaa. Generoiva jakauma on tasainen. Saadaan tulokset (5,1,6,4,3,1,1,6,2,4,1,3).



- Empiirinen jakauma *on* eräs tn-jakauma ja siinä voi käyttää tn-laskennan työkaluja.
- Mutta *eri jakauma* kuin generoiva.
- Jos n on suuri, jakaumat ovat lähellä toisiaan (SLL)

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)

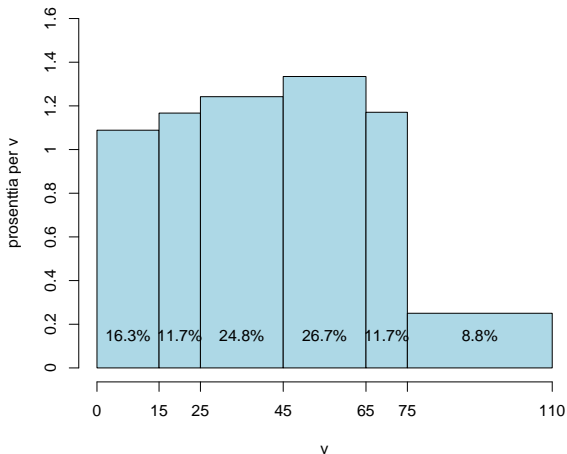
Esim: Suomalaiset

- 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 $((5,5), (1,2), (4,3), (4,3), (3,3), (5,5), (5,5))$

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.

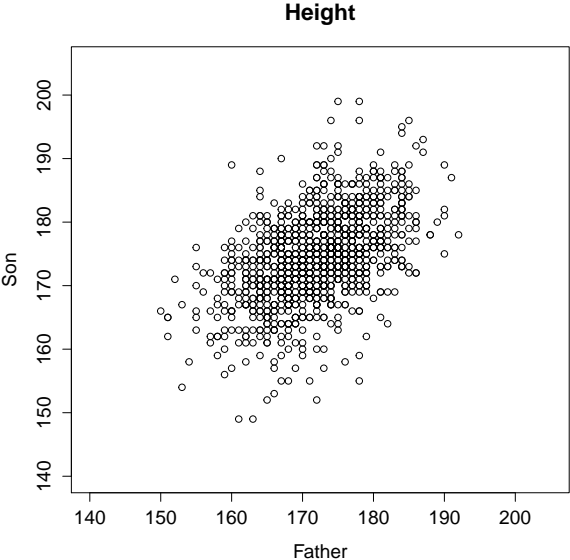
$$\text{cov}(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - m(\vec{x}))(y_i - m(\vec{y}))$$

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

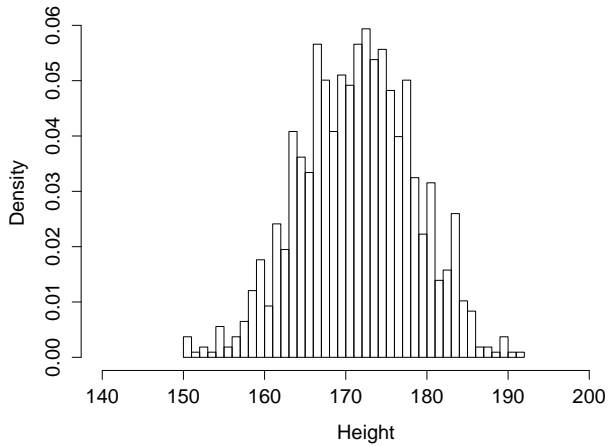
Esimerkki: Isien ja poikien pituudet

165	152	165	151	179	174	178	178	184	187	174	168	171	174	165	178	178	184	172	176	178	181	170	161	168	170	179	179	175	190	180	168	173	173									
161	161	166	171	181	172	177	178	173	182	173	168	169	172	169	178	181	183	175	165	177	170	175	175	178	182	158	167	170	178	155	157	172	173	170								
165	161	167	170	180	174	177	172	182	174	167	172	168	177	183	162	158	168	166	172	178	175	177	182	161	163	167	171	161	176	184	178	184	164	171	173							
167	159	167	171	182	173	180	178	185	196	181	169	169	173	167	177	184	182	160	169	160	174	175	175	177	182	166	164	167	170	163	176	186	170	151	165	163	170	173	174			
155	163	167	172	186	172	181	178	159	156	159	170	173	169	178	170	186	164	169	163	172	175	176	182	181	165	163	171	170	165	175	163	181	186	185	162	171	173	173				
160	163	167	171	166	175	183	179	189	160	170	172	169	179	172	187	162	168	162	172	175	176	183	181	169	164	169	171	176	166	174	183	186	161	174	173	174	173	173				
166	163	167	170	162	177	183	178	165	168	163	171	172	174	177	186	164	167	166	172	176	178	175	176	184	169	164	171	167	171	167	174	186	166	164	171	174	173					
164	163	170	172	166	177	192	178	187	191	162	172	173	174	178	177	186	164	173	177	176	184	182	151	165	171	170	168	175	170	180	168	155	164	171	176	173	173					
168	164	169	171	164	176	164	173	167	165	171	175	173	173	180	179	185	169	167	174	177	177	164	184	160	166	171	171	167	176	170	181	183	171	173	176	173	173					
170	163	171	170	169	177	169	180	159	163	166	171	176	172	173	178	177	185	168	168	172	177	170	170	184	163	167	173	170	176	174	182	173	180	170	159	176	174	174				
150	166	171	171	167	176	167	181	178	159	165	171	175	174	175	179	184	185	167	169	167	173	178	175	172	162	163	167	173	170	176	174	181	166	157	165	169	176	174				
160	166	169	171	167	176	170	181	187	174	175	174	175	179	185	186	170	168	168	172	181	177	173	183	164	165	173	170	170	175	173	182	179	171	176	167	178	171	173	173			
162	167	173	172	168	177	170	182	181	169	166	158	176	173	175	179	173	188	171	169	168	173	181	175	176	183	166	167	174	170	171	176	174	174	164	164	162	174	177	172			
163	166	172	171	169	176	173	181	176	158	163	163	177	173	176	178	174	190	171	168	171	173	182	176	175	184	168	167	175	171	173	176	174	181	157	161	164	172	179	170			
164	166	174	172	169	177	171	180	178	166	168	169	179	172	177	181	188	174	169	170	172	163	179	178	183	168	167	177	171	172	176	175	181	165	161	168	171	181	173				
165	165	172	171	169	175	173	181	166	163	175	167	178	174	177	178	183	189	172	167	170	173	166	178	178	183	170	165	178	172	174	176	175	182	168	162	168	171	183	174			
169	165	175	172	172	176	175	180	159	158	164	167	180	180	174	179	179	180	189	172	169	173	168	179	180	184	173	165	177	171	174	175	178	180	151	162	169	171	184	173			
167	166	175	171	174	176	176	181	162	161	168	170	179	173	178	179	191	191	174	168	171	174	167	178	181	184	174	167	178	172	174	176	177	182	159	162	168	170	160	177			
171	165	177	170	172	176	175	180	167	161	167	170	184	173	180	177	174	186	176	169	170	173	166	177	183	162	156	169	155	175	176	178	180	162	162	167	172	162	174				
170	166	179	169	172	177	177	181	173	160	169	171	184	174	181	179	153	154	155	170	173	174	169	179	185	184	161	168	159	174	175	175	178	181	164	163	170	170	164	175			
172	165	178	170	173	176	170	182	157	166	167	170	150	176	178	182	158	150	172	173	171	178	173	186	162	167	165	174	176	175	178	181	164	170	169	166	167	170	169	166			
177	166	185	172	176	175	178	181	162	162	169	171	162	175	182	177	160	169	163	171	172	173	171	178	173	178	175	178	186	163	168	163	173	175	176	179	181	168	164	169	170	167	175
159	169	160	173	174	176	178	180	165	164	171	172	164	176	186	176	187	149	163	170	173	173	173	177	175	186	165	168	165	174	175	175	184	181	169	162	170	171	168	175			
162	169	160	174	174	176	180	181	167	162	170	171	165	175	165	167	180	188	178	162	171	174	172	180	177	186	164	168	165	174	177	178	186	180	158	165	170	167	175				
164	168	162	173	174	176	181	182	169	164	169	172	166	176	168	181	183	180	164	170	174	174	173	178	181	187	168	168	168	172	177	176	164	185	166	172	171	169	175				
165	168	164	174	174	175	184	181	175	163	171	170	167	176	188	181	155	165	164	170	175	173	176	177	183	185	169	168	168	173	170	176	170	163	166	172	170	171	167	175			
164	168	164	174	179	176	186	180	167	166	171	170	168	175	171	180	166	179	173	173	175	173	175	179	185	185	169	167	169	173	170	175	172	184	165	166	171	170	175				
167	169	168	174	178	175	190	182	163	165	173	171	167	177	169	180	183	176	184	162	173	178	178	167	190	171	169	168	170	180	176	174	184	166	165	173	170	170	175				
166	167	167	173	178	176	171	184	162	167	172	171	170	176	173	181	165	168	160	174	173	174	177	174	188	171	169	167	174	181	176	176	184	166	166	174	170	174	177	177			
167	168	168	174	178	176	171	183	165	165	173	170	171	177	172	180	166	175	166	167	178	174	176	178	180	189	171	168	170	174	183	175	174	183	174	183	167	174	171	172	176		
170	167	166	173	181	177	172	184	166	165	173	170	171	175	173	180	165	177	177	171	177	178	178	184	189	188	172	169	171	174	161	177	182	168	164	176	172	173	166	175			
169	167	167	174	180	176	176	183	165	166	175	171	172	178	184	164	159	166	171	178	172	179	184	188	174	168	170	174	165	179	178	184	171	165	177	172	172	170	175				
169	169	170	173	183	176	175	184	167	167	176	170	172	175	176	181	163	161	166	172	182	174	178	179	175	192	172	167	169	174	186	167	178	179	184	172	167	178	171	172	175		
172	167	169	174	190	175	178	183	168	167	176	170	174	176	178	181	169	161	168	171	184	174	180	177	187	193	176	199	171	172	168	177	180	183	178	166	170	172	176	176			
173	169	170	173	165	177	177	184	170	165	179	170	172	177	182	181	170	166	167	180	183	174	180	177	178	199	177	169	170	173	169	178	182	184	161	169	160	172	175	177			
172	168	170	173	167	178	179	184	172	166	178	172	174	178	178	181	158	166	168	170	185	176	181	179	173	157	182	171	172	168	179	185	166	167	167	173	175	176	176				
174	168	170	174	167	178	181	184	175	164	179	170	174	176	177	182	163	168	168	170	160	176	184	178	164	177	158	171	174	170	177	173	183	166	168	164	174	175	177	177			
176	169	170	172	167	178	178	183	158	167	159	172	175	176	178	187	165	163	168	170	164	177	185	170	154	158	162	171	172	173	170	178	172	187	164	168	163	174	175	176			
182	169	172	173	167	179	184	183	159	160	172	174	175	180	181	167	164	170	170	166	176																						

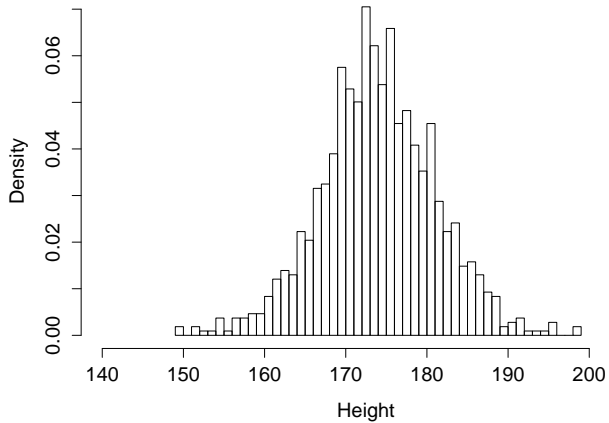
Hajontakuvio (scatterplot)



Histogram of Fathers



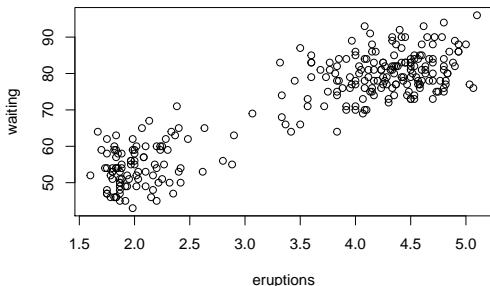
Histogram of Sons



Esimerkki: Old Faithful -geysirin purkaukset

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

2 muuttujaa: Purkauksen kesto ja väliaika seuraavaan purkaukseen.



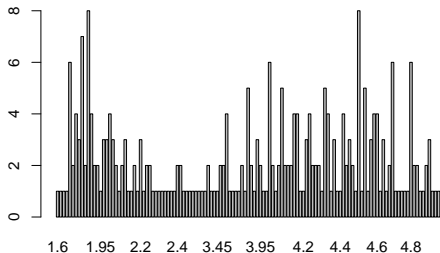
Datan silmäilystä (*eyeballing*) on hyvä aloittaa, jos se voi paljastaa ilmiön olennaisia piirteitä.

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

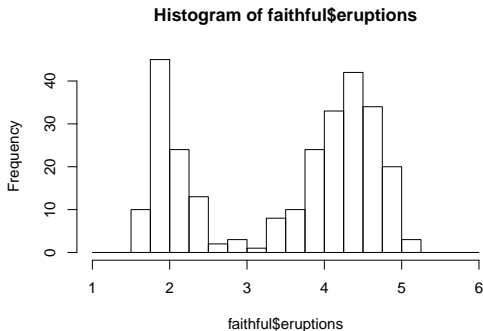
x	1.6	1.667	1.7	1.733	1.75	...	5.1
$n_{\bar{x}}(x)$	1	1	1	1	6	...	1



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)

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

on datajoukossa arvoltaan (x, y) olevien alkioiden lukumäärä.

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

	<hr/>					
	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	3	3
Yht	0	1	3	0	3	

Suhteellisten esiintyvyyksien ristitaulukko

Arvo x ja y suhteellinen esiintyvyys

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

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

	y					
x	1	2	3	4	5	Yht
1	0	$\frac{1}{7}$	0	0	0	$\frac{1}{7}$
2	0	0	0	0	0	0
3	0	0	$\frac{1}{7}$	0	0	$\frac{1}{7}$
4	0	0	$\frac{2}{7}$	0	0	$\frac{2}{7}$
5	0	0	0	0	$\frac{3}{7}$	$\frac{3}{7}$
Yht	0	$\frac{1}{7}$	$\frac{3}{7}$	0	$\frac{3}{7}$	

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

Empiirinen yhteisjakauma

Lause

Datajoukosta $\vec{x}\vec{y} = ((x_1, y_1), \dots, (x_n, y_n))$ tasaisen satunnaisesti valittu pari (X, Y) on diskreetti satunnaismuuttuja, joka noudattaa datajoukon $\vec{x}\vec{y}$ empiiristä jakaumaa tiheysfunktiona $f_{X,Y}(x, y) = f_{\vec{x}\vec{y}}(x, y)$ 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	mean()	np.mean()	AVERAGE()
$sd(\vec{x})$	Keskihajonta	sqrt(1-1/n)*sd()	np.std()	STDEV.P()
$sd_s(\vec{x})$	Otoskeskihajonta	sd()	np.std(,ddof=1)	STDEV.S()
$var(\vec{x})$	Varianssi	(1-1/n)*var()	np.var()	VAR.P()
$var_s(\vec{x})$	Otosvarianssi	var()	np.var(,ddof=1)	VAR.S()
$cov(\vec{x}, \vec{y})$	Kovarianssi	(1-1/n)*cov()	np.cov(,ddof=0)[0][1]	COVARIANCE.P()
$cov_s(\vec{x}, \vec{y})$	Otoskovarianssi	cov()	np.cov(,ddof=1)[0][1]	COVARIANCE.S()
$cor(\vec{x}, \vec{y})$	Korrelaatio	cor()	np.corrcoef()[0][1]	CORREL()
$q_{0.5}(\vec{x})$	Mediaani	median()	np.median()	MEDIAN()
$q_{0.25}(\vec{x})$	Alakvartiili	quantile(, .25)	np.quantile(, .25)	PERCENTILE.INC(, .25)
$q_{0.75}(\vec{x})$	Yläkvartiili	quantile(, .75)	np.quantile(, .75)	PERCENTILE.INC(, .75)

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

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% havainnoista

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

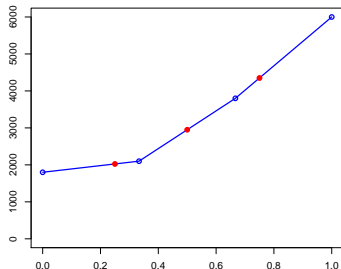
Äärellisellä datalla vain “suunnilleen”, koska esim. 9:stä havainnosta ei voi ottaa tasan puolta. Tähän on erilaisia ratkaisuja, esim. seuraavaksi esitettävä kvantiilifunktio.

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

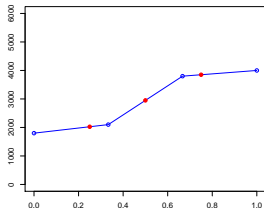
Esim. Piirrä seuraavien datajoukkojen kvantiilifunktiot ja määritä niiden mediaanit ja keskiarvot:

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

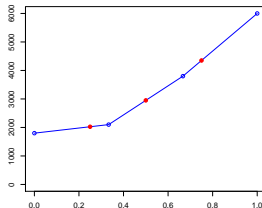
$$\vec{y} = (1800, 2100, 3800, 6000)$$

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

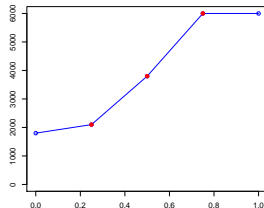
(Huom. Viimeisen joukon koko ei ole neljällä jaollinen.)



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



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



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

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 nopanheittotulosta	Potentiaalinen ääretön heittojono

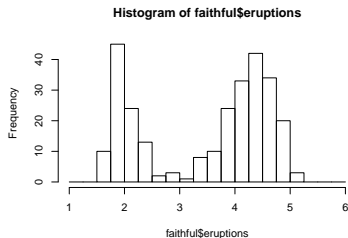
Populaatio on tilastotieteen terminologiaa, ja tarkoittaa

- 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* 272 purkauspituutta. Fysikaalinen mekanismi on ehkä monimutkainen, mutta ajatellaan pituuksien käyttäytyvän **kuten** satunnaismuuttuja, jolla on eräs **jakauma** f . Mutta mikä jakauma?



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

Empiirinen jakauma **apksimoi** 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 . . .