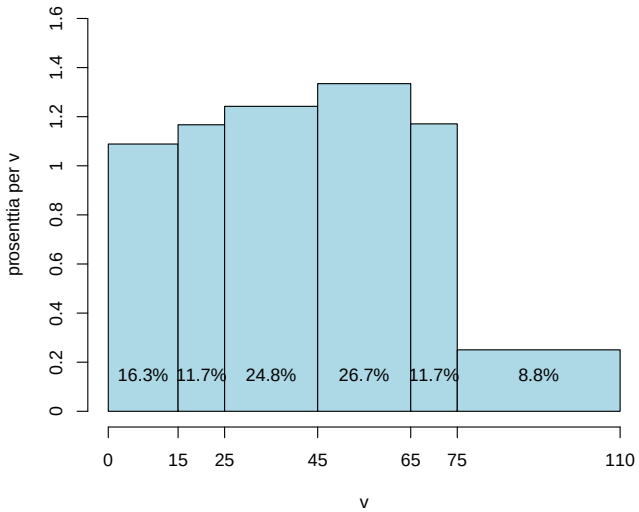
4A Datajoukkojen kuvaajat ja tunnusluvut

Tuntitehtävät

4A1 (Ikäjakauma) Seuraava taulukko ja histogrammi kuvaavat Suomen väestörakennetta ikäluokittain 31.12.2015. Ikää pidetään tässä tehtävässä reaalilukuna, esim. ikä 14.9 kuuluu puoliavoimeen väliin $[0, 15)$.

Ikäluokka (v)	Lukumäärä
$[0, 15)$	896 023
$[15, 25)$	640 387
$[25, 45)$	1 363 155
$[45, 65)$	1 464 640
$[65, 75)$	642 428
$[75, 110]$	480 675

(Lähde: Tilastokeskus)



Vastaamaan seuraaviin kysymyksiin luokitellun datan perusteella. Kohdissa (a)–(c) oleta, että iät jakautuvat kunkin luokan sisällä tasaisesti.

(a) Kumpia on väestössä enemmän, 1-vuotiaita vai 66-vuotiaita? (1-vuotiaalla tarkoitetaan henkilöä, jonka ikä reaalilukuna on välillä $[1, 2)$.)

(b) Mikä on väestön mediaani-ikä (ikä, jonka alapuolella on puolet väestöstä)?

(c) Mikä on väestön iän keskiarvo?

(d) Mitä pystyt sanomaan iän mediaanista ja keskiarvosta, mikäli oletusta tasaisesta jakaumasta kunkin luokan sisällä ei voida tehdä? Voitko laskea ne täsmälleen? Jos et, etsi kummallekin luvulle pienin ja suurin mahdollinen arvo.

4A2 (Kvantiilit) Datajoukon $x = (x_1, \ldots, x_n)$ kvantiilifunktio määritetään R:ssä oletusarvoisesti seuraavasti. Merkitään $x_{(1)}$ = datajoukon pienin arvo, $x_{(2)}$ = toiseksi pienin, jne. Tällöin data saadaan järjestettyä muotoon $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Seuraavaksi jaetaan vaaka-akselin yksikköväli $n-1$ yhtä pitkään osaväliin käyttämällä välien reunapisteinä lukuja $p_k = (k-1)/(n-1)$, $k = 1, \ldots, n$. Kvantiilifunktion $Q(p)$ kuvaaja piirretään piirtämällä (x, y)-tasoon ensin pisteet $(p_k, x_{(k)})$, $k = 1, \ldots, n$, ja sen jälkeen yhdistämällä pisteet suorilla viivoilla.

Piirrä (käsin) paperille seuraavien datajoukkojen kvantiilifunktiot ja määritä kuvaajasta näille alakvartiili $Q(0.25)$, mediaani $Q(0.50)$ ja yläkvartiili $Q(0.75)$:

(a) $x = (1000, 2000, 5000, 9000)$,

(b) $x = (1000, 2000, 2000, 8000, 9000)$,

(c) $x = (1, 20, 1, 5, 1)$.

Tutki seuraavaksi seuraavia väitteitä. Perustele miksi väite on tosi tai kehittele vastaesimerkki, jonka perusteella väite on epätosi.

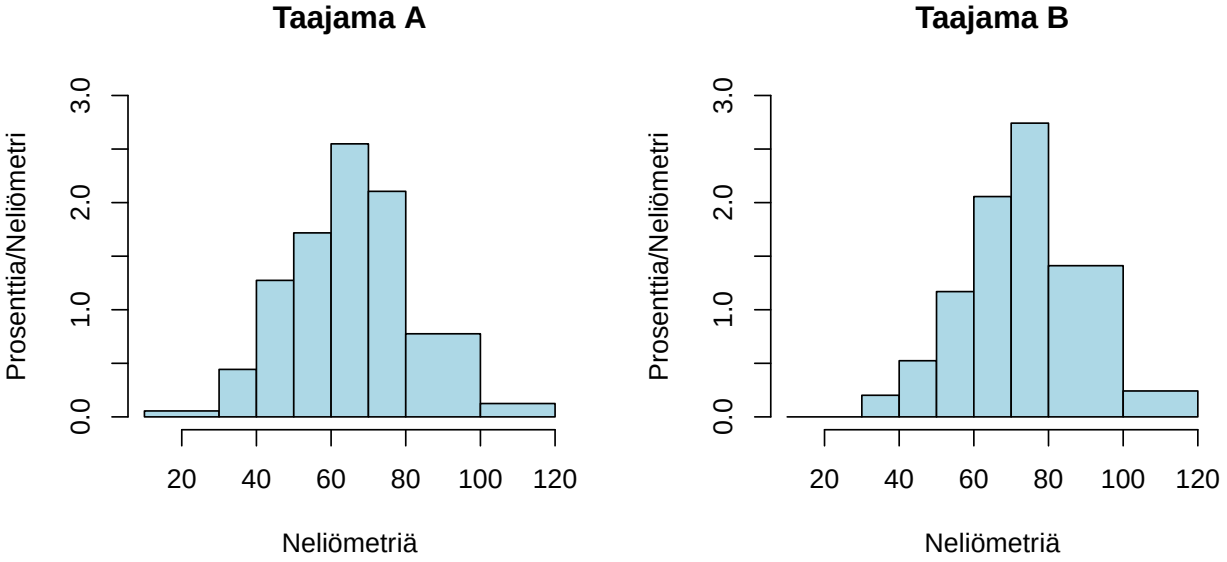
(d) Datajoukon keskiarvo on aina yhtäsuuri kuin sen mediaani.

(e) Datajoukon alakvartiili on aina pienempi tai yhtä kuin mediaani.

(f) Datajoukon alakvartiili on aina pienempi tai yhtä suuri kuin keskiarvo.

Kotitehtävät

4A3 (Asuinneliöt.) Taajamassa A on 361 asuntoa ja taajamassa B on 248 asuntoa. Alla olevissa histogrammeissa on esitettynä kummankin taajaman asuntojen pinta-alat.



Vastaa histogrammien avulla seuraaviin kysymyksiin (oletetaan yksinkertaisuuden vuoksi, että yksikään havaituista pinta-aloista ei osu luokkien reunoille).

(a) Kuinka monessa taajaman B asunnossa pinta-ala on vähintään 80 m^2?

(b) Kumman taajaman asuntojen pinta-alan mediaani on suurempi? Pystytäänkö tähän kysymykseen vastaamaan, jos pinta-alojen ei oleteta jakautuvan luokkien sisällä tasaisesti?

4A4 (Kaksi noppaa) Luennoija suoritti $n = 18$ kertaa kokeen, jossa heitettiin punaista ja keltaista noppaa. Tulokset muodostavat datajoukon $((r_1, y_1), \ldots, (r_n, y_n))$. Seuraava ristitaulukko esittää kaikkien mahdollisten arvoparien (r, y) esiintyvyydet (lukumäärät) tässä datajoukossa. Suhteelliset esiintyvyydet saa tietysti jakamalla luvut n:llä.

		r					
		1	2	3	4	5	6
	1	0	0	0	0	0	0
	2	0	0	1	2	0	0
y	3	0	0	0	1	0	0
	4	1	0	1	0	0	0
	5	1	0	0	0	0	0
	6	2	4	1	1	2	1

(a) Määritä kummankin nopan tulosten $(r_1, \ldots, r_n)$ ja $(y_1, \ldots, y_n)$ empiiriset jakaumat (kaksi eri jakaumaa). Laske kummankin datajoukon keskiarvo (eli empiirisen jakauman odotusarvo).

(b) Laske kummankin datajoukon keskihajonnat.

(c) Laske kahden muuttujan empiirisen jakauman korrelaatiokerroin. Vihje: Laske ensin empiirisen jakauman mukainen $\mathrm{E}(RY)$ ristitaulukon avulla. Käytä sitten kaavaa $\mathrm{Cov}(R, Y) = \mathrm{E}(RY) - \mathrm{E}(R)\,\mathrm{E}(Y)$. Lopuksi laske korrelaatiokerroin kovarianssista.

(d) Onko empiirisestä jakaumasta laskemasi korrelaatiokerroin negatiivinen, nolla vai positiivinen? Kuvaile sanallisesti, mitä tämä kertoo datajoukosta.

(e) Yllä tarkasteltiin heittotulosten empiiristä jakaumaa. Siirrytään nyt pohtimaan sitä stokastista prosessia (generoivaa jakaumaa), josta heittotulokset ovat peräisin. Jos R ja Y ovat satunnaismuuttujat, jotka kuvaavat punaisen ja keltaisen nopan heittoa, arveletko arkijärjen ja yllä tehtyjen havaintojen perusteella, että R ja Y ovat riippuvat vai riippumattomat? Entä millainen korrelaatiokerroin niillä on *generoivassa jakaumassa*?

(f) Arvioi empiiristen jakaumien perusteella, ovatko punainen ja keltainen noppa reiluja. (Tämä ei ole laskutehtävä, vaan päättely- ja arviointitehtävä.)

Vihje: Empiirisiä jakaumia ja ristitaulukoita on käsitelty luennolla 3B.