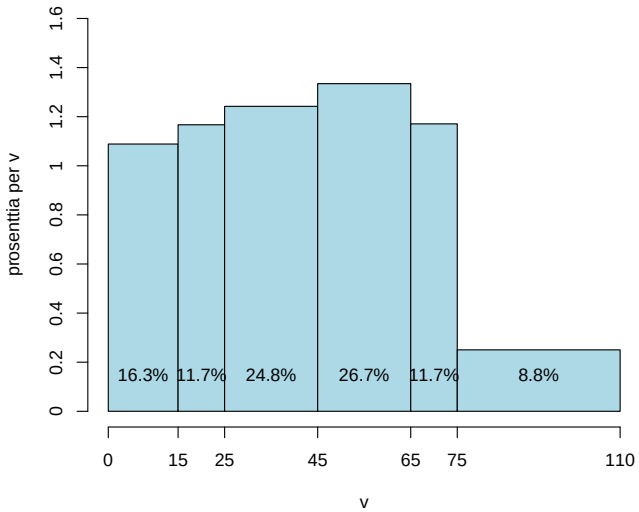
4A Datajoukkojen kuvaajat ja tunnusluvut

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.

Ratkaisu.

(a) Luokassa $[0, 15)$ on 16.3% väestöstä, jolloin tasajakaumaoletuksen mukaan välille $[1, 2)$ osuu noin $16.3\%/15 \approx 1.09\%$ väestöstä. Vastaavasti välille $[66, 67)$ osuu noin $11.7\%/10 \approx 1.17\%$ väestöstä (huom. luokkavälien pituudet erisuuret). Näin arvioituna 66-vuotiaita on hiukan enemmän kuin 1-vuotiaita.

(b) Kahdessa ensimmäisessä luokassa eli välillä $[0, 25)$ on 28.0% väestöstä. Jotta saadaan puolet väestöstä, tähän tarvitaan vielä lisää 22.0% väestöstä eli osuus $22.0/24.8 \approx 0.8871$

kolmannesta luokasta $[25, 45)$. Koska kolmas luokkaväli on 20 vuotta pitkä, tasajakaumaoletuksen mukaan haluttu osuus ihmisiä löytyy kolmannesta luokasta pisteen $25 + 0.8871 \cdot 20 \approx 25 + 17.7 = 42.7$ alapuolelta. Mediaani on siis noin 42.7 vuotta.

(c) Tasajakaumaoletuksen mukaan kunkin luokan sisällä ikien keskiarvo on luokkavälin keskellä, esim. ensimmäisessä luokassa $(0+15)/2 = 7.5$ vuotta ja toisessa luokassa $(15+25)/2 = 20$ vuotta. Koko väestön iän keskiarvo on ikien summa jaettuna väestön koolla,

$$\frac{896023 \cdot 7.5 + 640387 \cdot 20 + 1363155 \cdot 35 + 1464640 \cdot 55 + 642428 \cdot 70 + 480675 \cdot 92.5}{5487308}$$

eli noin 43.2 vuotta. Tämä on luokkien keskikohtien *painotettu keskiarvo*, jossa painot ovat luokkien koot (ihmismäärät). Sama tulos olisi luonnollisesti saatu, jos painoina olisi käytetty luokkien osuuksia väestöstä (prosentteina).

(d) Mediaani on varmasti välillä $[25, 45]$. Mitä keskiarvoon tulee, jos kussakin luokassa kaikki iät olisivat luokkavälin alarajalla, niin väestön iän keskiarvoksi tulisikin

$$\frac{896023 \cdot 0 + 640387 \cdot 15 + 1363155 \cdot 25 + 1464640 \cdot 45 + 642428 \cdot 65 + 480675 \cdot 75}{5487308}$$

eli noin 34.2 vuotta. Jos taas iät olisivat luokkavälien ylärajalla, keskiarvoksi tulisi noin 52.3 vuotta. Väestön iän keskiarvo on siis jossain välillä $[34.2, 52.3]$.

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.

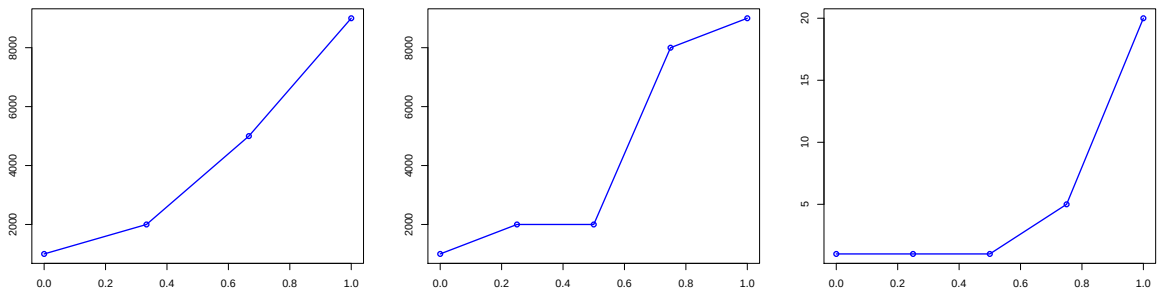
(f) Datajoukon alakvartiili on aina pienempi tai yhtä suuri kuin keskiarvo.

Ratkaisu.

(a) Kuvaaja alla.
Alakvartiili $Q(0.25) = 1750$, mediaani $Q(0.50) = 3500$, yläkvartiili $Q(0.75) = 6000$.

(b) Kuvaaja alla.
Alakvartiili $Q(0.25) = 2000$, mediaani $Q(0.50) = 2000$, yläkvartiili $Q(0.75) = 8000$.

(c) Kuvaaja alla.
Alakvartiili $Q(0.25) = 1$, mediaani $Q(0.50) = 1$, yläkvartiili $Q(0.75) = 5$.



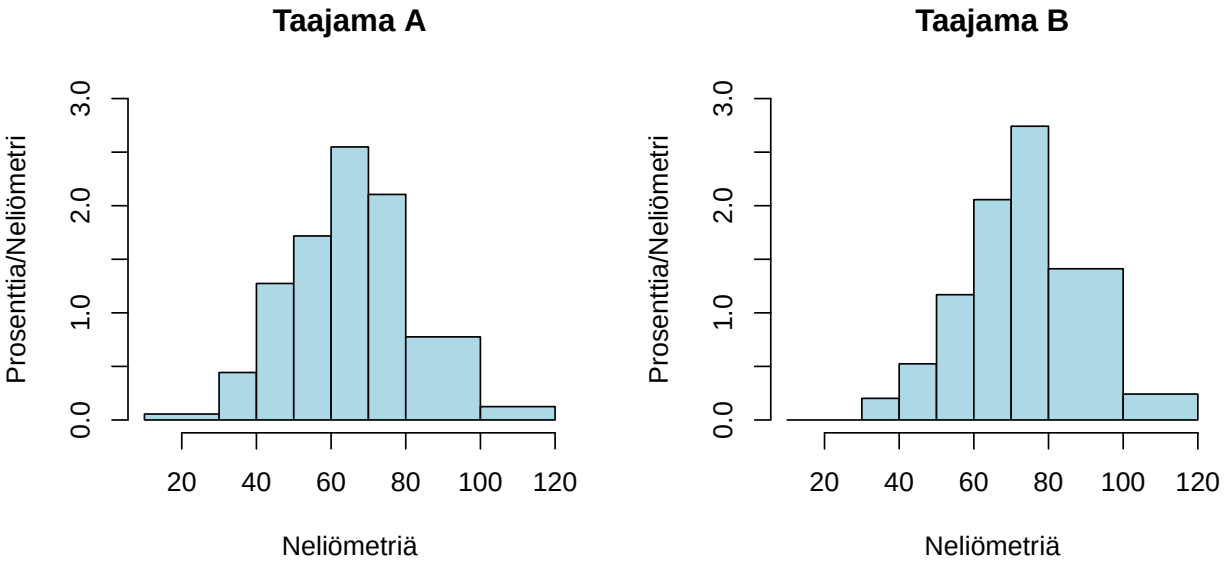
(d) Ei pidä paikkaansa. Vastaesimerkiksi kelpaa mikä vain kohdista (a)–(c).

(e) Tämä on totta, sillä kvantiilifunktio on ei-vähenevä.

(f) Ei pidä paikkaansa. Yksi erittäin pieni poikkeava arvo voi siirtää keskiarvon valtavan pieneksi, jopa alakvartiilin alapuolelle. Esimerkiksi datajoukon $x = (-1000, 1, 1, 5, 20)$ kvartiilit ovat samat kuin (c)-kohdassa, mutta keskiarvo -194.6 on huimasti alle alakvartiilin.

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?

Ratkaisu.

(a) Taajamassa B palkkien korkeudet luokkaväleillä 80–100 ja 100–120 ovat likimain 1.4 ja 0.25. Kertomalla nämä luokkavälien pituuksilla saadaan vähintään 80-neliöisten asuntojen osuudeksi likimain

$$1.4 \times 20 + 0.25 \times 20 \;=\; 33\%.$$

Koska taajamassa B on yhteensä 248 asuntoa, on niistä vähintään 80-neliöisiä siis likimain $248 \times 0.33 \approx 82$ kappaletta.

(b) Lähdetään jakaumien yläpäästä liikkeelle ja selvitetään kuinka monta palkkia täytyy kulkea ennen kuin saadaan katettua vähintään 50% kaikista havainnoista. Palkki, jossa 50% ylittyy sisältää pinta-alojen mediaanin.

Taajamassa B kolmen oikeanpuoleisimman palkin korkeudet ovat noin $2.75, 1.4, 0.25$, jolloin vastaaviin luokkiin kuuluvien asuntojen osuudet saadaan kertomlla korkeudet luokkien leveyksillä. Tuloksiksi saadaan likimain $27.5\%, 28\%, 5\%$. Näiden summa on yli 50%,

joten taajamassa B neliöiden mediaani on siis vähintään 70 (tarkalleen ottaen mediaani on välillä 70–80).

Vastaavat laskutoimitukset taajaman A kolmelle oikeanpuoleisimmalle palkille antavat likimain korkeudet $2.1, 0.75, 0.1$ ja osuudet $21\%, 15\%, 2\%$. Nyt osuuksien summa on 38%, siis alle 50%, ja mediaanin täytyy olla siis alle 70.

Taajaman B neliöiden mediaani on siis suurempi kuin taajaman A, ja tämä voidaan sanoa vaikka ei tunneta (tai oleteta) jakaumaa tasaiseksi kunkin luokan sisällä.

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.

Ratkaisu.

(a) Punaisen nopan empiirinen jakauma (ristitaulukon sarakesummat jaettuna 18:lla):

r	1	2	3	4	5	6
$f_{\vec{r}}(r)$	4/18	4/18	3/18	4/18	2/18	1/18

Keskiarvo on $m(\vec{r}) = \mathbf{2.9444}$.

Keltaisen nopan empiirinen jakauma (ristitaulukon rivisummat jaettuna 18:lla):

y	1	2	3	4	5	6
$f_{\vec{y}}(y)$	0	3/18	1/18	2/18	1/18	11/18

Keskiarvo on $m(\vec{y}) = \mathbf{4.8889}$.

(b)
$$\mathrm{sd}(\vec{r}) = \sqrt{\sum_{r=1}^{6}(r - m(\vec{r}))^2 f_{\vec{r}}(r)} = \mathbf{1.5082}$$

$$\mathrm{sd}(\vec{y}) = \sqrt{\sum_{y=1}^{6}(y - m(\vec{y}))^2 f_{\vec{y}}(y)} = \mathbf{1.5595}$$

Laskuissa voi myös käyttää vaihtoehtoista laskukaavaa $\mathrm{Var}(R) = \mathrm{E}(R^2) - (\mathrm{E}(R))^2$.

(c) Vihjeen mukaisesti lasketaan

$$\mathrm{E}(RY) = \sum_{r=1}^{6}\sum_{y=1}^{6} ry f_{\vec{r}\vec{y}}(r, y) = 14.0556.$$

(Vaikka arvoparilla (r, y) on 36 mahdollista arvoa, vain 12 niistä havaittiin datajoukossa, joten summassa on vain 12 nollasta poikkeavaa termiä.)

Sitten

$$\mathrm{Cov}(R, Y) = \mathrm{E}(RY) - \mathrm{E}(R)\,\mathrm{E}(Y) = 14.0556 - 2.9444 \cdot 4.8889 = -0.3393$$

(huom. tässä käytetään todennäköisyyslaskennan mukaisia merkintöjä, mutta koko ajan *empiirisessä* jakaumassa). Lopuksi

$$\mathrm{Cor}(R, Y) = \frac{\mathrm{Cov}(R, Y)}{\mathrm{SD}(R)\,\mathrm{SD}(Y)} = \mathbf{-0.1442}.$$

(d) Empiirinen korrelaatiokerroin on negatiivinen. Tämä kertoo siitä, että havaitussa datajoukossa keltaisen nopan saadessa suuria arvoja punainen noppa on saanut pieniä arvoja (ks. erityisesti taulukon alin rivi).

(e) Fysikaalisesti ajatellen vaikuttaa todennäköiseltä, että R ja Y ovat **riippumattomat** (keltaisen nopan arvoilla on aina sama jakauma, riippumatta siitä mikä tulos punaisesta nopasta tuli). Siispä niiden korrelaatiokerroinkin lienee **nolla**.

Eräs tehtävän opetuksista on, että empiirinen jakauma yleensä poikkeaa jonkin verran generoivasta jakaumasta, ja niin poikkeavat sen tunnusluvutkin kuten keskiarvo ja korrelaatiokerroin.

(f) Kummankaan nopan empiirinen jakauma ei ole täsmälleen tasajakauma. Punaisen nopan empiirinen jakauma on kuitenkin *melko lähellä* tasajakaumaa, joten punainen noppa voi hyvinkin olla (ainakin melkein) reilu.

Keltaisen nopan empiirinen jakauma poikkeaa hyvin paljon tasajakaumasta. Vaikka datajoukko onkin pieni (18 heittoa), tuntuisi varsin yllättävältä, että tällaiset tulokset olisi saatu reilusta nopasta. Vaikuttaa siltä, että keltainen noppa on epäreilu, erityisesti se tuottaa kuutosia paljon suuremmalla todennäköisyydellä kuin 1/6.

Myöhemmin kurssilla tullaan oppimaan menetelmiä, joilla nämä arvelut saadaan matemaattisempaan muotoon.