J Kohonen

3A Keskihajonta ja korrelaatio

Tuntitehtävät

3A1 (Korrelaatio ja riippuvuus) Diskreettien satunnaismuuttujien X ja Y yhteisjakauma on esitetty allaolevana taulukkona:

	Y		
X	-1	0	1
-1	0	$\frac{1}{6}$	$\frac{1}{6}$
0	$\frac{1}{3}$	0	0
1	0	$\frac{1}{6}$	$\frac{1}{6}$

(a) Määritä X:n jakauma, odotusarvo ja keskihajonta.

(b) Määritä Y:n jakauma, odotusarvo ja keskihajonta.

(c) Laske X:n ja Y:n korrelaatio.

(d) Selvitä, ovatko X ja Y riippuvat vai riippumattomat.

Ratkaisu.

(a) Yhteisjakauman taulukon rivisummat laskemalla saadaan X:n jakaumaksi

k	-1	0	1
$\mathrm{P}(X=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Näin ollen X:n odotusarvoksi saadaan

$$\mathrm{E}(X) = \sum_k k\,\mathrm{P}(X=k) = (-1)\times\frac{1}{3} + 0\times\frac{1}{3} + 1\times\frac{1}{3} = 0.$$

Keskihajonta saadaan varianssin neliöjuurena, ja varianssi puolestaan on ehkä kätevintä laskea kaavalla $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$. Koska

$$\mathrm{E}(X^2) = \sum_k k^2\,\mathrm{P}(X=k) = (-1)^2\times\frac{1}{3} + 0^2\times\frac{1}{3} + 1^2\times\frac{1}{3} = \frac{2}{3},$$

havaitaan, että X:n keskihajonta on

$$\mathrm{SD}(X) = \sqrt{\frac{2}{3} - 0^2} = \sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2} = \sqrt{\frac{2}{3}} \approx 0.82.$$

(b) Yhteisjakauman taulukon sarakesummat laskemalla saadaan Y:n jakaumaksi

k	-1	0	1
$\mathrm{P}(Y=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Tästä nähdään, että X ja Y ovat samoin jakautuneita. Näin ollen Y:n odotusarvo ja keskihajonta ovat samat mitä X:lläkin, eli $\mathrm{E}(Y)=0$ ja $\mathrm{SD}(Y)=\sqrt{2/3}\approx 0.82$.

(c) Korrelaatio saadaan normittamalla kovarianssi, ja kovarianssi yleensä saadaan helpoiten laskukaavasta $\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)$. Yhteisjakauman taulukon avulla tulee laskea

$$\mathrm{E}(XY) \;=\; \sum_i \sum_j ij\,\mathrm{P}(X=i, Y=j).$$

Ylläolevasta summasta voidaan jättää pois kaikki termit, joissa $i=0$ tai $j=0$ (koska niiden kontribuutio summaan on nolla), joten

$$\begin{aligned}\mathrm{E}(XY) &= \sum_{i\neq 0}\sum_{j\neq 0} ij\,\mathrm{P}(X=i,Y=j)\\ &= (-1)\times(-1)\times 0 \;+\; (-1)\times 1\times\frac{1}{6} \;+\; 1\times(-1)\times 0 \;+\; 1\times 1\times\frac{1}{6}\\ &= 0.\end{aligned}$$

Näin ollen X:n ja Y:n korrelaatio on

$$\mathrm{Cor}(X,Y) \;=\; \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; \frac{\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; 0.$$

(d) X ja Y ovat stokastisesti riippuvat, sillä esimerkiksi

$$\mathrm{P}(Y=-1\,|\,X=0) \;=\; 1,$$

kun taas

$$\mathrm{P}(Y=-1) \;=\; 1/3$$

Tämän tehtävän viesti on, että korreloimattomuus *ei* takaa riippumattomuutta. Stokastinen riippumattomuus on vahva ominaisuus (kertolaskukaavan pitäisi päteä joka kohdassa yhteisjakaumaa; tai toisin sanottuna, Y:n ehdollisten jakaumien pitäisi olla samat jokaisella arvolla $X=x$). Stokastisesta riippumattomuudesta seuraa kyllä korreloitumattomuus.

Korrelaatio puolestaan mittaa vain tietyntyyppistä, luonteeltaan lineaarista, stokastista riippuvuutta. Muuttujilla voi olla epälineaarinen stokastinen riippuvuus, kuten tässä tehtävässä, ja niiden korrelaatio voi silti olla nolla.

3A2 (Nopanheittojen keskiarvo) Tavallista noppaa heitetään monta kertaa peräkkäin. Heittojen tuloksia merkitään $X_1, X_2, \ldots$ ja ne ovat keskenään riippumattomat. Ensimmäisten n:n tuloksen keskiarvoa merkitään $A_n = \frac{1}{n}(X_1+\ldots+X_n)$.

(a) Laske satunnaismuuttujan X_1 odotusarvo ja keskihajonta.

(b) Määritä satunnaismuuttujan $A_2 = \frac{1}{2}(X_1 + X_2)$ jakauma.
Vihje: Selvitä ensin A_2:n arvojoukko. Tutki sitten, aluksi pienillä arvoilla a, milloin eli millä parin (X_1, X_2) arvoilla toteutuu $A_2 = a$. Koeta yleistää päättelysi.

(c) Laske satunnaismuuttujan A_2 odotusarvo ja keskihajonta. Vertaa muuttujien A_2 ja X_1 keskihajontoja laskemalla niiden suhde.

(d) Laske odotusarvo ja keskihajonta satunnaismuuttujalle

$$A_{100} = \frac{1}{100}(X_1 + X_2 + \cdots + X_{100}).$$

Vertaa muuttujien A_{100} ja X_1 keskihajontoja laskemalla niiden suhde.

Opastus. C-kohdan voit laskea joko b-kohdassa lasketusta jakaumasta, tai voit hyödyntää odotusarvon ja varianssin yhteenlaskuominaisuuksia. D-kohdassa tarkan jakauman selvittäminen olisi hyvin työlästä, joten on parempi käyttää yhteenlaskuominaisuuksia.

Lauseesta 4.9 seuraa, että $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X, Y)$. **Riippumattomilla** satunnaismuuttujilla em. kaavasta jää kovarianssitermi nollaksi, ja usealle **riippumattomalle** muuttujalle kaava yleistyy muotoon $\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$. Summien variansseihin tutustutaan tarkemmin luentomonisteen luvussa 5.

Huomaa, että yhteenlaskukaava pätee nimenomaan variansseille, ei keskihajonnoille.

Ratkaisu.

(a) X_1 noudattaa joukon $\{1, 2, \ldots, 6\}$ tasajakaumaa, joten sen odotusarvo saadaan kaavasta

$$\mathrm{E}(X_1) = \sum_k k\,\mathrm{P}(X_1 = k) = \sum_{k=1}^{6} k \times \frac{1}{6} = \frac{7}{2} = \mathbf{3.5}.$$

Keskihajonnan laskemiseksi lasketaan ensin X_1:n neliön odotusarvo

$$\mathrm{E}(X_1^2) = \sum_k k^2\,\mathrm{P}(X_1 = k) = \sum_{k=1}^{6} k^2 \times \frac{1}{6} = \frac{91}{6}.$$

Tämän jälkeen keskihajonta saadaan kaavasta

$$\mathrm{SD}(X_1) = \sqrt{\mathrm{E}(X_1^2) - \mathrm{E}(X_1)^2} = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{35}{12}} \approx \mathbf{1.7078}.$$

(b) Tarkastelemalla mitä arvoja X_1 ja X_2 voivat saada, havaitaan että A_2:n mahdollisten arvojen joukko on $\{1.0, 1.5, 2.0, 2.5, \ldots, 5.5, 6.0\}$. Kyseisten arvojen todennäköisyydet voidaan määrittää kohta kohdalta:

- Tapahtuma $\{A_2 = 1\}$ sattuu täsmälleen silloin, kun $X_1 = 1$ ja $X_2 = 1$. Näin ollen $\mathrm{P}(A_2 = 1) = (\frac{1}{6})^2 = \frac{1}{36}$.

- Tapahtuma $\{A_2 = 1.5\}$ sattuu, jos $(X_1, X_2) = (1, 2)$ tai $(X_1, X_2) = (2, 1)$. Näin ollen $\mathrm{P}(A_2 = 1.5) = 2 \times (\frac{1}{6})^2 = \frac{2}{36}$.
- ...

Näin etenemällä saadaan A_2:n jakauma määritettyä taulukkoon:

k	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
$\mathrm{P}(A_2 = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

(c) **Tapa 1.** Satunnaismuuttujan A_2 odotusarvo ja keskihajonta voidaan määrittää b)-kohdan taulukosta kaavoilla:

$$\mathrm{E}(A_2) \;=\; \sum_x x\,\mathrm{P}(A_2 = x) \;=\; 1.0 \times \frac{1}{36} + 1.5 \times \frac{2}{36} + \cdots \;=\; 3.5$$

ja

$$\mathrm{E}(A_2^2) \;=\; \sum_x x^2\,\mathrm{P}(A_2 = x) \;=\; 1.0^2 \times \frac{1}{36} + 1.5^2 \times \frac{2}{36} + \cdots \;\approx\; 13.71,$$

josta

$$\mathrm{SD}(A_2) \;=\; \sqrt{\mathrm{E}(A_2^2) - (\mathrm{E}(A_2))^2} \;\approx\; \sqrt{13.71 - 3.5^2} \;\approx\; 1.2.$$

Lukuarvot voi laskea taskulaskimella tai esim. R-komennoilla

```
x <- seq(1.0,6.0,by=0.5)
p <- c(1:6,5:1)/36
m1 <- sum(x*p)
m2 <- sum(x^2*p)
mu <- m1
sigma <- sqrt(m2-m1^2)
```

Tapa 2. Koska X_2:llä on sama jakauma kuin X_1:llä, sillä on myös sama odotusarvo ja varianssi, jotka laskettiin a-kohdassa. Odotusarvon laskusäännöillä

$$\mathrm{E}(A_2) = \mathrm{E}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{2}(\mathrm{E}(X) + \mathrm{E}(Y)) = \frac{1}{2}(3.5 + 3.5) = \mathbf{3.5}.$$

Varianssin laskusäännöillä (vrt. opastus)

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2 \cdot \mathrm{Cov}(X_1, X_2) = (35/12) + (35/12) + 2 \cdot 0 = 35/6,$$

missä kovarianssi oli nolla, koska X_1 ja X_2 ovat riippumattomat. Lasketaan vielä keskihajonta ja lopuksi skaalataan kertoimella puoli.

$$\mathrm{SD}(X_1 + X_2) = \sqrt{35/6}$$

ja

$$\mathrm{SD}(A_2) = \mathrm{SD}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{2}\mathrm{SD}(X_1 + X_2) = \frac{1}{2}\sqrt{35/6} = \sqrt{35/24} \approx \mathbf{1.2076}.$$

Lasketaan vielä kysytty suhde.

$$\frac{\mathrm{SD}(A_2)}{\mathrm{SD}(X_1)} = \frac{\sqrt{35/24}}{\sqrt{35/12}} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}}} \approx \mathbf{0.7071}.$$

(d) Satunnaismuuttujan A_{100} arvojoukko on $\{1.00, 1.01, 1.02, \ldots, 5.99, 6.00\}$. Tämä 501 rationaaliluvun arvojoukko voitaisiin periaatteessa listata taulukkoon ja laskea jokaisen lukuarvon todennäköisyys. Ei ole kuitenkaan aivan yksinkertaista laskea esim. tapahtuman $\{A_{100} = 3.97\}$ todennäköisyyttä käymällä läpi kaikkia 100 nopan kombinaatioita, jotka tuottavat keskiarvoksi 3.97, sillä sadan nopan tuloskombinaatioita on tähtitieteellisen suuri määrä: $6^{100} \approx 6 \cdot 10^{77}$ kappaletta.

Satunnaismuuttujan A_{100} odotusarvo voidaan helposti laskea lineaarisuutta käyttämällä:

$$\mathrm{E}(A_{100}) \;=\; \mathrm{E}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) \;=\; \frac{1}{100}\sum_{i=1}^{100}\mathrm{E}(X_i) \;=\; \frac{1}{100}\sum_{i=1}^{100} 3.5 \;=\; \mathbf{3.5}.$$

(Summan termit olivat samat, koska kaikki noppatulokset $X_1, X_2, \ldots$ noudattavat samaa lukujoukon $\{1, \ldots, 6\}$ tasajakaumaa.)

Satunnaismuuttujan A_{100} varianssi voidaan laskea yleisten varianssin laskusääntöjen avulla tuntematta A_{100}:n jakaumaa, sillä tulokset $X_1, X_2, \ldots$ ovat toisistaan stokastisesti riippumattomat. Lisäksi tiedetään että jokaisella heittotuloksella X_i on sama jakauma ja siis sama varianssi kuin tuloksella X_1. Näin ollen

$$\mathrm{Var}(A_{100}) \;=\; \mathrm{Var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) \;=\; \left(\frac{1}{100}\right)^2 \sum_{i=1}^{100}\mathrm{Var}(X_i) \;=\; \frac{\mathrm{Var}(X_1)}{100}$$

ja A_{100}:n keskihajonnaksi saadaan (a)-kohdan avulla

$$\mathrm{SD}(A_{100}) \;=\; \sqrt{\frac{\mathrm{Var}(X_1)}{100}} \;=\; \frac{\mathrm{SD}(X_1)}{10} \;=\; \frac{\sqrt{35/12}}{10} \;\approx\; \mathbf{0.1708}.$$

Nähdään, että kysytty suhde on

$$\frac{\mathrm{SD}(A_{100})}{\mathrm{SD}(X_1)} = \mathbf{0.10}$$

eli sadan heittotuloksen keskiarvon keskihajonta on **kymmenesosa** yhden heittotuloksen keskihajonnasta.

Tehtävä osoittaa, että riippumattomien ja samoin jakautuneiden satunnaismuuttujien keskiarvoistaminen *säilyttää odotusarvon ennallaan* mutta *pienentää keskihajontaa*. Tämä tärkeä havainto selittää, miten riskejä voi pienentää hajauttamalla. Tähän myös pohjautuu suurten finanssi- ja vakuutusyhtiöiden sekä kasinojen toiminta. Toisaalta tähän perustuu myös tilastotieteellinen estimoiminen, koska otoksen suurentaminen saa otoksen keskiarvon osumaan lähemmäs jakauman odotusarvoa, eli tieto odotusarvosta tarkentuu.

Kotitehtävät

3A3 (Lämpötilamalli) Meteorologi mallintaa tämän ja huomisen päivän lämpötilojen T_0 ja T_1 välistä yhteyttä kaavalla

$$T_1 \;=\; T_0 + \Delta T$$

jossa ΔT kuvaa lämpötilojen muutosta. Satunnaismuuttujat T_0 ja ΔT oletetaan toisistaan riippumattomiksi. Lisäksi tiedetään, että $\mathrm{E}(T_0) = \mu$ ja $\mathrm{Var}(T_0) = \sigma^2$ sekä $\mathrm{E}(\Delta T) = 0$ ja $\mathrm{Var}(\Delta T) = \theta^2$. Mallin parametrit μ, σ ja θ oletetaan ennalta tunnetuiksi, ja lisäksi $\sigma > 0$ ja $\theta \geq 0$.

(a) Määritä $\mathrm{E}(T_1)$.

(b) Määritä $\mathrm{SD}(T_1)$. Vrt. tehtävään 3A2. Miten lasketaan summan varianssi?

(c) Määritä $\mathrm{Cov}(T_1, T_0)$. Käytä kovarianssin bilineaarisuutta.

(d) Määritä $\mathrm{Cor}(T_1, T_0)$. Ennen kuin lasket korrelaation tarkan arvon, yritä intuitiivisesti päätellä korrelaation tulkinnan kautta miten se käyttäytyy tapauksissa $\theta = 0$ ja $\theta \gg \sigma$ (eli θ paljon suurempi kuin σ).

Tulokset tulee ilmoittaa mahdollisimman yksinkertaisina lausekkeina mallin parametreista μ, σ ja θ.

Arviointiohje. 0.5 pistettä per kohta. Tehtävän kokonaispistemäärä pyöristetään ylöspäin lähimpään kokonaislukuun. Jos väärin laskettua arvoa käytetään myöhemmässä kohdassa oikealla tavalla, ei tästä sakoteta enää uudelleen.

Ratkaisu.

(a) Odotusarvon lineaarisuuden perusteella

$$\mathrm{E}(T_1) = \mathrm{E}(T_0 + \Delta T) = \mathrm{E}(T_0) + \mathrm{E}(\Delta T) = \mu + 0 = \mu.$$

Siis huomisen lämpötilan odotusarvo on sama kuin tämän päivän lämpötilan odotusarvo.

(b) Satunnaismuuttujien summan varianssin kaavan perusteella

$$\begin{aligned}
\mathrm{Var}(T_1) &= \mathrm{Var}(T_0 + \Delta T) \\
&= \mathrm{Var}(T_0) + 2\,\mathrm{Cov}(T_0, \Delta T) + \mathrm{Var}(\Delta T) \\
&= \sigma^2 + 0 + \theta^2 \\
&= \sigma^2 + \theta^2,
\end{aligned}$$

missä $\mathrm{Cov}(T_0, \Delta T) = 0$ koska T_0 ja ΔT ovat riippumattomia. Siis $\mathrm{SD}(T_1) = \sqrt{\sigma^2 + \theta^2}$.

(c) Käyttäen kovarianssin laskusääntöjä havaitaan, että

$$\begin{aligned}\operatorname{Cov}(T_1, T_0) &= \operatorname{Cov}(T_0 + \Delta T,\, T_0) \\ &= \operatorname{Cov}(T_0, T_0) + \operatorname{Cov}(\Delta T, T_0) \\ &= \operatorname{Var}(T_0) + 0 \\ &= \sigma^2.\end{aligned}$$

(d) Korrelaatio mittaa sitä, kuinka hyvin satunnaismuuttujalla voi (lineaarisesti) ennustaa toisen satunnaismuuttujan arvoja. Muutoksen keskihajonnan θ ollessa lähellä nollaa, pitäisi huomisen lämpötilan olla paremmin ennustettavissa tämän päivän lämpötilasta. Muutoksen keskihajonnan θ kasvattaminen suureksi puolestaan tarkoittaa, että huomisen lämpötila on keskimäärin kauempana tämän päivän lämpötilasta, ja täten heikommin ennustettava. Näin ollen on oletettavaa, että korrelaatio lähestyy kohti nollaa muutoksen keskihajonnan θ kasvaessa suureksi. Varmistetaan tämä laskemalla, kovarianssin bilineaarisuuden avulla:

$$\operatorname{Cor}(T_1, T_0) = \frac{\operatorname{Cov}(T_1, T_0)}{\operatorname{SD}(T_1)\operatorname{SD}(T_0)} = \frac{\sigma^2}{(\sqrt{\sigma^2 + \theta^2}) \cdot \sigma} = \frac{\sigma}{\sqrt{\sigma^2 + \theta^2}}.$$

Viimeinen lauseke on tosiaan θ:n suhteen vähenevä funktio, vahvistaen aiemman intuition. Pisteessä $\theta = 0$ lauseke antaa korrelaation arvoksi täsmälleen $+1$, mikä vastaa tarkkaa lineaarista riippuvuutta. Jos taas θ on hyvin suuri, niin lausekkeen arvo on lähellä nollaa.

3A4 (Kustannusfunktion minimointi) Eräs pieni kaupunki sijoittuu itä-länsi-suuntaisen tien varteen niin, että mallinnamme kaupungin 2 kilometrin pituisena reaalilukuvälinä $[0, 2]$. Kaupungin itäosassa asukastiheys on suurempi; jos valitsemme satunnaisen asukkaan, niin hänen asuinpaikkansa X on jatkuva satunnaismuuttuja, jonka tiheysfunktio on $f(x) = x/2$, kun $0 \le x \le 2$. (Todellisuudessa populaatio olisi äärellinen ja asuinpaikka diskreetti, mutta oletamme asukkaita olevan niin paljon, että voimme käsitellä heitä jatkuvana massana.)

(a) Laske asuinpaikan odotusarvo $\mu = \operatorname{E}(X)$. Laske asuinpaikan mediaani, eli sellainen piste m että $\operatorname{P}(X \le m) = \frac{1}{2}$, ts. puolet asukkaista asuu pisteen länsipuolella.

(b) Laske $\operatorname{E}(X^2)$ ja $\operatorname{SD}(X)$.

(c) Abel on kaupunkisuunnittelija, jonka tehtävänä on valita kaupungintalolle paikka a johonkin kohtaa kaupunkia ($0 \le a \le 2$). Hän pyrkii minimoimaan *neliöllisen kustannusfunktion* $q(a) = \operatorname{E}\left((X - a)^2\right)$. Toisin sanoen, hän haluaa minimoida kaupungin asukkaiden *keskimääräisen neliöidyn etäisyyden* asuinpaikastaan kaupungintalolle.

Määritä q yksinkertaisena a:stä riippuvana (polynomi)lausekkeena, jossa ei ole satunnaismuuttujia eikä integraaleja. Mikä on tämän funktion $q(a)$ muoto? Muokkaa annettua $q(a)$:n lauseketta sellaiseen muotoon, että pääset käyttämään odotusarvon lineaarisuutta. (Vaihtoehtoisesti voit käyttää odotusarvon muunnoskaavaa ja laskea näin saamasi integraalin.) Funktion muotoa voi tarkastella lausekkeesta tai piirtämällä kuvan.

Etsi se arvo a, jolla $q(a)$ on mahdollisimman pieni. (Minimikohdan pitäisi löytyä helposti derivaatan avulla.) Onko se jompikumpi arvoista μ ja m?

(d) Bertta on toinen suunnittelija. Hänen tehtävänään on valita kirjastolle paikka b johonkin kohtaan kaupunkia ($0 \le b \le 2$). Hän pyrkii minimoimaan *lineaarisen kustannusfunktion* $\ell(b) = \mathrm{E}(\,|X - b|\,)$. Toisin sanoen, hän haluaa minimoida kaupungin asukkaiden *keskimääräisen etäisyyden* asuinpaikastaan kirjastoon.

Määritä $\ell(b)$ yksinkertaisena b:stä riippuvana (polynomi)lausekkeena. Aloita odotusarvon muunnoskaavasta. Integraali kannattaa sitten jakaa kahteen osaan, eli väleihin $[0, b]$ ja $(b, 2]$, jotta pääset eroon itseisarvomerkeistä.

Etsi se arvo b, jolla $\ell(b)$ on mahdollisimman pieni. Onko se jompikumpi arvoista μ ja m?

(e) Vapaaehtoinen ylimääräinen lisätehtävä (ei pisteitä): Ovatko Abelin ja Bertan valitsemat paikat samat, vai onko jompikumpi idempänä? Pohdi miksi. Voiko ilmiön ymmärtää intuitiivisesti tarkastelemalla heidän kustannusfunktioitaan? Tapahtuisiko sama ilmiö, vaikka asukkaiden jakauma olisi jokin muu kuin tässä tehtävässä annettu? Entä jos kaupunki ei olekaan jana, vaan alue tasossa?

Arviointiohje. Kohdista a–d 0.5 pistettä per kohta, summa pyöristetään ylöspäin. Oikeasta laskusta saa pisteet vaikka laskussa käyttäisi aiemmasta kohdasta tullutta väärää numeroarvoa.

Ratkaisu.

(a)

$$\mu = \mathrm{E}(X) = \int_0^2 x f(x)dx = \int_0^2 \frac{x^2}{2}dx = \Big/_0^2 \left(\frac{x^3}{6}\right) = \frac{8}{6} = \frac{4}{3} \approx 1.333.$$

Mediaanin löytämiseksi todetaan ensin, että

$$\mathrm{P}(X \le m) = \int_0^m f(x)dx = \int_0^m \frac{x}{2}dx = \Big/_0^m \left(\frac{x^2}{4}\right) = \frac{m^2}{4}.$$

Ehdosta $m^2/4 = \frac{1}{2}$ ratkaistaan helposti $m = \sqrt{2} \approx 1.414$.

(b)

$$\mathrm{E}(X^2) = \int_0^2 x^2 f(x)dx = \int_0^2 \frac{x^3}{2}dx = \Big/_0^2 \left(\frac{x^4}{8}\right) = \frac{16}{8} = 2.$$

Tästä voidaan edelleen laskea

$$\mathrm{SD}(X) = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2} = \sqrt{2 - (4/3)^2} = \sqrt{2/9} \approx 0.471.$$

(c) Jos Abel sijoittaa kaupungintalon paikkaan a, niin paikassa x asuvan kaupunkilaisen neliöity etäisyys kaupungintalolle on $(x-a)^2$. Neliöllinen kustannusfunktio on

$$q(a) = \mathrm{E}((X-a)^2) = \mathrm{E}(X^2 - 2aX + a^2) = \mathrm{E}(X^2) - 2a\,\mathrm{E}(X) + a^2 = 2 - \frac{8}{3}a + a^2,$$

missä vihjeen mukaisesti käytettiin odotusarvon lineaarisuutta. Nyt $q(a)$ on ylöspäin aukeava paraabeli, ja sen minimikohta löytyy derivaatan

$$q'(a) = 2a - \frac{8}{3}$$

nollakohdasta. Asettamalla derivaatan nollaksi saamme ratkaisun $a = \frac{4}{3} \approx 1.333$. Huomaamme, että Abel sijoittaa kaupungintalon kohtaan μ eli asukkaiden keskimääräiseen asuinpaikkaan.

(d) Jos Bertta sijoittaa kirjaston paikkaan b, niin paikassa x asuvan kaupunkilaisen etäisyys kirjastoon on $|x - b|$, toisin sanoen $x - b$ jos hän asuu kirjaston itäpuolella ($x \geq b$), ja $b - x$ jos hän asuu kirjaston länsipuolella ($x < b$).

Lineaarinen kustannusfunktio on

$$\begin{aligned}
\ell(b) = \mathrm{E}\left(|X - b| \right) &= \int_0^2 |x - b| \, f(x) dx \\
&= \int_0^b (b - x) f(x) dx \; + \int_b^2 (x - b) f(x) dx \\
&= \int_0^b \left(\frac{bx}{2} - \frac{x^2}{2} \right) dx + \int_b^2 \left(\frac{x^2}{2} - \frac{bx}{2} \right) dx \\
&= \frac{b^3}{6} - b + \frac{4}{3}.
\end{aligned}$$

Funktio on kolmannen asteen polynomi. Piirtämällä, tai tarkastelemalla ensimmäistä ja toista derivaattaa, toteamme että $\ell(b)$ kaartuu ylöspäin kaikkialla välillä $(0, 2)$, joten sen minimi löytyy derivaatan nollakohdasta. Derivaatta on

$$\ell'(b) = \frac{1}{2} b^2 - 1,$$

joka saa arvon nolla kun $b = \sqrt{2} \approx 1.414$. Huomaamme, että Bertta sijoittaa kirjaston paikkaan m, eli asukkaiden mediaaniasuinpaikkaan (ts. puolet asukkaista asuu kirjaston länsipuolella ja puolet itäpuolella).

(e) Sekä Abel että Bertta valitsevat paikan kaupungin itäosasta, koska siellä asuu enemmän ihmisiä. Abelin *neliöllinen* kustannusfunktio antaa suuremman painoarvon sille, että joku asuu hyvin kaukana kaupungintalosta (koska etäisyys neliöidään). Välttääkseen hyvin suuria etäisyyksiä kaupungin länsiosan asukkaisiin, Abel ei vie kaupungintaloa aivan niin itään kuin Bertta kirjaston.

Sama ilmiö havaittaisiin muillakin asukasjakaumilla. Yleisemmin, voidaan todistaa että odotusarvo on se kohta jossa neliöllinen kustannusfunktio minimoituu, ja mediaani on se kohta jossa lineaarinen kustannusfunktio minimoituu. Voit kokeilla todistaa nämä yleiset väittämät (todistus ei ole paljonkaan vaikeampi kuin tässä tehtävässä tehty).

Jos kaupunki olisi tasoalue, neliöllisen kustannusfunktion minimointi onnistuisi sekin melko helposti, mutta lineaarisen kustannusfunktion minimointi olisi hiukan hankalampaa (googlaa “geometric median”).