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.

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

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

(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:

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

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) Abel ja Bertta ovat töissä eri sääennusteyrityksissä. He ovat kumpikin simuloineet tietokoneella mahdollisia säätiloja ja laskeneet ennusteen huomisen keskipäivän lämpötilalle X celsiusasteina. Kumpikin on päätynyt samaan tulokseen, että lämpötila on satunnaismuuttuja, jolla on kolmion muotoinen tiheysfunktio $f(x) = x/18$ välillä $[0, 6]$ (ja nolla muualla).

Kummankaan työnantaja ja sääennusteita odottavat asiakkaat eivät kuitenkaan halua kuulla mistään "jakaumista", vaan he haluavat yksinkertaisesti *piste-ennusteen* eli yhden lämpötilaluvun. Abelin on siis valittava ennusteekseen jokin luku $a \in [0, 6]$, samoin Bertan on valittava jokin luku $b \in [0, 6]$. He voivat valita eri luvut, jos haluavat.

(a) Tarkista integroimalla, että f tosiaan on kelvollinen jatkuvan jakauman tiheysfunktio.

(b) Laske X:n *odotusarvo* $\mu = \mathrm{E}(X)$ sekä *mediaani* eli sellainen luku m, että $\mathrm{P}(X \le m) = \frac{1}{2}$.

(c) Abelin työnantaja kannustaa häntä osumaan lähelle oikeaa siten, että hänen palkastaan vähennetään *neliöllinen sakko* $(X-a)^2$, jossain sopivissa rahayksiköissä, kun X on huomenna havaittu lämpötila ja a on Abelin ennustama lämpötila. Koska X:n arvo ei ole vielä tiedossa, Abel pyrkii valitsemaan a:n siten, että hänen sakkonsa odotusarvo olisi mahdollisimman pieni, ts. hän haluaa minimoida funktion $q(a) = \mathrm{E}\left((X-a)^2\right)$. Sievennä funktio q sellaiseen muotoon, että se on yksinkertainen a:n polynomi, jossa ei ole E-merkkejä.

Vihje: Voit esim. aloittaa kertomalla binomin neliön auki. Vaihtoehtoisesti voit aloittaa huomaamalla, että $(X-a)^2$ on eräs X:n muunnos, ja esittää sen odotusarvon integraalina muunnoksen odotusarvon kaavalla (luento 2A).

Piirrä $q(a)$ välillä $a \in [0, 6]$ sen muodon hahmottamiseksi. Abel valitsee ennusteensa a siten, että $q(a)$ on mahdollisimman pieni. Etsi tämä luku. Onko se jompikumpi luvuista μ tai m?

(d) Myös Bertan työnantaja kannustaa osumaan lähelle oikeaa, mutta käyttää *lineaarista sakkoa*, $|X - b|$, jossain rahayksiköissä, kun X on havaittu lämpötila ja b on hänen ennusteensa. Bertan sakon odotusarvo on $\ell(b) = \mathrm{E}(\,|X - b|\,)$. Kirjoita tämä integraalina ja sievennä se yksinkertaiseksi b:n polynomiksi. Piirrä funktio ja selvitä, minkä luvun Bertta valitsee, jotta $\ell(b)$ minimoituu. Onko se jompikumpi luvuista μ tai m?

Vihje: Integraali kannattaa hajottaa kahteen osaan, integraaliksi välillä $x \in [0, b)$ ja integraaliksi välillä $x \in [b, 6]$, jolloin pääset eroon itseisarvomerkeistä. Huom: Bertan ongelma johtaa vähän hankalampaan integraaliin kuin Abelin, mutta laske integraalit sinnikkäästi auki.

(e) Jos Abel ja Bertta antoivat eri lämpötilaennusteet, miksi? Koeta selittää arkijärjellä, miten eri sakkofunktiot vaikuttivat heidän toimintaansa.

(f) (Vapaaehtoinen lisätehtävä, vaikeampi) Jos sait selville, että Abelin ja Bertan ennusteet vastaavat ennustejakauman tiettyjä kohtia, koeta todistaa että sama pätisi vaikka ennustejakauma olisi *mikä tahansa* tiheysfunktio.

Tämä on esimerkki sijoitteluongelmasta (engl. *facility location problem*, `https://en.wikipedia.org/wiki/Facility_location_problem`). Sama matemaattinen muoto tulee vastaan esim. valittaessa jonkin palvelun (kauppa, kirjasto, tukkuvarasto) sijaintia, jos halutaan minimoida esim. käyttäjien keskimääräinen etäisyys kyseiseen palveluun, ja käyttäjien sijaintien jakauma tunnetaan. Keskimääräisen etäisyyden minimointi vastaa Bertan ongelmaa; keskimääräisen *neliöidyn* etäisyyden minimointi vastaa Abelin ongelmaa.

Ratkaisu.

(a) Funktio on kaikkialla epänegatiivinen. Tarkistetaan että lisäksi sen integraali on ykkönen:

$$\int_0^6 f(x)dx = \int_0^6 (x/18)\,dx = \Big/_{x=0}^{6} \left(\frac{x^2}{36}\right) = 36/36 = 1.$$

(b)

$$\mu = \mathrm{E}(X) = \int_0^6 x\ f(x)dx = \int_0^6 (x^2/18)dx = \Big/_{x=0}^{6} \left(x^3/54\right) = 4.$$

Mediaanin etsiminen:

$$\mathrm{P}(X \le m) = \int_0^m f(x)dx = \int_0^m (x/18)\,dx = \Big/_{x=0}^{m} \left(x^2/36\right) = m^2/36.$$

Merkitsemällä ko. lauseke yhtäsuureksi kuin $\frac{1}{2}$ ja ratkaisemalla m saadaan $m = \sqrt{18} \approx 4.2426$.

(c) Seurataan ensimmäistä vihjettä. Jos Abelin ennuste on a, niin sakon odotusarvo on

$$q(a) = \mathrm{E}((X - a)^2) = \mathrm{E}(X^2 - 2aX + a^2) = \mathrm{E}(X^2) - 2a\,\mathrm{E}(X) + a^2. \qquad (*)$$

Tiedämme jo luvun $\mathrm{E}(X) = 4$, ja tarvitsemme vielä luvun

$$\mathrm{E}(X^2) = \int_0^6 x^2 \, f(x) \, dx = \int_0^6 (x^3/18) \, dx = \Big/_{x=0}^{6} \left(() \, x^4/72 \right) = 18.$$

Siispä

$$q(a) = 18 - 2a \cdot 4 + a^2 = a^2 - 8a + 18.$$

Huomataan, että q on ylöspäin kaartuva paraabeli (2. derivaatta on positiivinen) ja sen minimikohta löytyy derivaatan

$$q'(a) = 2a - 8$$

nollakohdasta, joka on $a = 4$. Huomaamme, että Abelin valitsema ennuste on lämpötilan odotusarvo.

Huomataan lisäksi, että koska $a = \mu$, niin Abelin sakon odotusarvo on itse asiassa $\mathrm{E}((X-a)^2) = \mathrm{E}((X-\mu)^2) = \mathrm{Var}(X)$.

(d) Jos Bertan ennuste on b, niin hänen sakkonsa odotusarvo on

$$\begin{aligned}
\ell(b) = \mathrm{E}\left(|X - b| \right) &= \int_0^6 |x - b| \; f(x) dx \\
&= \int_0^b (b - x) f(x) dx \; + \; \int_b^6 (x - b) f(x) dx \\
&= \int_0^b (bx/18 - x^2/18) \, dx + \int_b^6 (x^2/18 - bx/18) \, dx \\
&= \Big/_0^b \left(bx^2/36 - x^3/54 \right) + \Big/_b^6 \left(x^3/54 - bx^2/36 \right) \\
&= b^3/54 - b + 4.
\end{aligned}$$

Tämä on kolmannen asteen polynomi. Tarkastelemalla sen muotoa (kuvasta tai 2. ja 3. derivaatasta) huomaamme, että $\ell(b)$ kaartuu ylöspäin kaikkialla välillä $[0, 6]$. Sen derivaatalla

$$\ell'(b) = b^2/18 - 1$$

on kyseisellä välillä nollakohta $b = \sqrt{18} \approx 4.2426$, ja tämä on siis funktion minimikohta. Bertta valitsee ennusteekseen lämpötilan mediaanin m.

(e) Sekä Abel että Bertta pyrkivät välttämään suuria ennustevirheitä, mutta koska Abelin ennusteen virhe *neliöidään*, hänellä on vielä Berttaakin voimakkaampi motiivi niiden välttämiseen. Jos Abel asettaisi ennusteensa esim. lähelle välin loppupistettä (6 astetta), ja lämpötila sattuisikin lähelle välin alkupistettä (0 astetta), niin Abelin sakko olisi aika suuri. Välttääkseen tällaisen suuren sakon mahdollisuuden hän asettaa ennusteensa hiukan alemmaksi kuin Bertta.

Toisin sanoen neliöllinen sakko johtaa varomaan suuria virheitä vielä voimakkaammin kuin lineaarinen sakko. Kumpi sitten on "parempi" sakkofunktio, riippuu täysin siitä millaisia ennusteita yhtiö haluaa. Sakkofunktioita voisi muotoilla muunkinlaisia, esim. sen perusteella, millaiset todelliset kustannukset syntyvät minkäkinlaisesta virheestä.

(f) (Vapaaehtoinen lisätehtävä) Sama ilmiö tosiaan tapahtuisi muillakin jakaumilla: jos halutaan minimoida neliöllinen sakko, kannattaa valita odotusarvo, ja jos halutaan minimoida lineaarinen sakko, kannattaa valita mediaani.

Odotusarvon tapaus on yllättävän helppo todistaa. Huomataan, että yllä yhtälössä (*) funktiossa $q(a)$ on kolme termiä, mutta ensimmäinen termi $\mathrm{E}(X^2)$ ei mitenkään riipu a:sta, joten se katoaa derivoitaessa. Funktion derivaatta onkin

$$q'(a) = 0 - 2\,\mathrm{E}(X) + 2a$$

ja tämä on nolla täsmälleen silloin kun $a = \mathrm{E}(X)$. Tässä ei mitenkään käytetty tietoa siitä, mikä jakauma X:llä on, joten tulos pätee kaikilla jakaumilla.

Mediaanin kohdalla todistus on hiukan pidempi (löytyy esim. googlaamalla “median minimizes expected absolute deviation”).