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.

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.

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.

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 $\mathrm{P}(X \leq m) = \frac{1}{2}$.

(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 odutusarvo 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.