

MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

2B Keskihajonta ja korrelaatio

Jukka Kohonen

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2022–2023
Periodi II

Sisältö

Varianssi ja keskihajonta

Poikkeaman todennäköisyyden yläraja

Kovarianssi ja korrelaatio

Mitä odotusarvo kertoo jakaumasta?

Satunnaismuuttujan odotusarvo $\mathbb{E}(X)$:

- on X :n mahdollisten arvojen todennäköisyyksillä painotettu summa, $\sum_x x f(x)$ tai $\int x f(x) dx$
- kertoo likiarvon keskiarvolle, joka saadaan suuresta määrästä riippumattomia X :n tavoin jakautuneita satunnaislukuja
- ei kerro mitään jakauman **leveystä**

Esim

Diskreettejä satunnaismuuttujia, joilla sama odotusarvo 1:

k	1
$\mathbb{P}(X = k)$	1

k	0	1	2
$\mathbb{P}(Z = k)$	$\frac{1}{2}$	0	$\frac{1}{2}$

k	0	1	2
$\mathbb{P}(Y = k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

k	0	1000000
$\mathbb{P}(W = k)$	0.999999	0.000001

Miten mitata satunnaismuuttujan poikkeamaa odotusarvosta? (Ensimmäinen yritys)

Satunnaisluvun (itseis)poikkeama odotusarvosta $\mu = \mathbb{E}(X)$ on satunnaisluku $|X - \mu|$. Sekin on satunnaisluku.

Jos nopanheitossa ($\mu = 3.5$) sattuu $X = 2$, niin $X - \mu = -1.5$.

Poikkeaman odotusarvo $\mathbb{E}|X - \mu|$:

- esim. nopalle $\frac{1}{6}(2.5 + 1.5 + 0.5 + 0.5 + 1.5 + 2.5) = 1.5$.
- kertoo likiarvon keskiarvolle $\frac{1}{n} \sum_{i=1}^n |X_i - \mu|$ suuresta määrästä riippumattomia X :n tavoin jakautuneita satunnaislukuja
- on optimoinnin kannalta hankala suure, koska funktio $x \mapsto |x|$ ei ole derivoituva nollassa.
- (ja eräitä muita matemaattisia hankaluuksia)

Entä jos korvataan $|X - \mu|$ luvulla $(X - \mu)^2$?

Varianssi (engl. *variance*)

Satunnaisluvun **neliöpoikkeama** odotusarvosta $\mu = \mathbb{E}(X)$ on satunnaisluku $(X - \mu)^2$. Sekin on satunnaisluku.

Esim. jos nopanheitossa ($\mu = 3.5$) sattuu $X = 2$, niin silloin neliöpoikkeama on $(2 - 3.5)^2 = (-1.5)^2 = 2.25$.

Neliöpoikkeaman odotusarvo eli satunnaisluvun X **varianssi** $\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$:

- esim. nopassa
 $\frac{1}{6}(2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2) \approx 2.917$
- kertoo likiarvon keskiarvolle $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ suuresta määrästä riippumattomia X :n tavoin jakautuneita sat.lukuja
- on optimoinnin kannalta mukava suure, koska funktio $x \mapsto x^2$ on äärettömän monta kertaa derivoituva
- noudattaa mukavaa yhteenlaskukaavaa (toisin kuin itseispoikkeama)

Varianssin tulkinta

Varianssi on yksiköltään neliö-jotain

	X	$\text{Var}(X)$
Pituus	m	m^2
Aika	s	s^2
Tuotto	EUR	EUR^2

Tulos palautetaan alkuperäisiin mittayksiköihin ottamalla neliöjuuri.
Varianssin neliöjuurta kutsutaan **keskihajonnaksi**, merk. $\text{SD}(X)$.

Esim. nopanheitto tuloksen keskihajonta

$$\sqrt{\frac{1}{6}(2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2)} \approx \sqrt{2.917} \approx 1.708.$$

(Vertaa itseispoikkeaman odotusarvoon 1.5.)

Keskihajonta (engl. *standard deviation*)

Satunnaisluvun keskihajonta $SD(X) = \sqrt{\mathbb{E}[(X - \mu)^2]}$ on alkuperäisiin yksiköihin normitettu odotusarvoinen neliöpoikkeama odotusarvosta $\mu = \mathbb{E}(X)$. Myös muita merkintöjä kuten $\mathbb{D}(X)$.

$SD(X)$ mittaa:

- paljonko X yleensä poikkeaa odotusarvostaan (hiukan kiertotietä: neliöpoikkeaman odotusarvon neliöjuurena)
- X :n jakauman leveyttä

Diskreetti jakauma:

$$\mu = \sum_x x f(x)$$

$$SD(X) = \sqrt{\sum_x (x - \mu)^2 f(x)}$$

Jatkuva jakauma:

$$\mu = \int x f(x) dx$$

$$SD(X) = \sqrt{\int (x - \mu)^2 f(x) dx}$$

Esimerkki: Erilaisia satunnaislukuja, joilla odotusarvo 1

Laske satunnaislukujen X , Y , Z keskihajonnat:

k	1
$\mathbb{P}(X = k)$	1

k	0	1	2
$\mathbb{P}(Y = k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

k	0	2
$\mathbb{P}(Z = k)$	$\frac{1}{2}$	$\frac{1}{2}$

$$\text{SD}(X) = \sqrt{\sum_k (k - \mu)^2 f_X(k)} = \sqrt{(1 - 1)^2 \times 1} = 0.$$

$$\text{SD}(Y) = \sqrt{(0 - 1)^2 \times \frac{1}{3} + (1 - 1)^2 \times \frac{1}{3} + (2 - 1)^2 \times \frac{1}{3}} = \sqrt{\frac{2}{3}} \approx 0.82.$$

$$\text{SD}(Z) = \sqrt{(0 - 1)^2 \times \frac{1}{2} + (1 - 1)^2 \times 0 + (2 - 1)^2 \times \frac{1}{2}} = 1.$$

Keskihajonta: Vaihtoehtoinen laskentakaava

Fakta

Jos satunnaisluvun X odotusarvo on $\mu = \mathbb{E}(X)$, niin pätee

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}(X^2) - \mu^2}.$$

(Tämä on joskus laskuissa kätevämpi, jos $\mathbb{E}(X^2)$ on helppo laskea. Siihen voi käyttää odotusarvon muunnoskaavaa.)

Todistus.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2\end{aligned}$$

$$\implies \text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[X^2] - \mu^2}$$



Esimerkki: Musta joutsen — Kahden arvon jakauma

k	0	10^6
$\mathbb{P}(X = k)$	$1 - 10^{-6}$	10^{-6}

$$\mu = \mathbb{E}(X) = 1$$

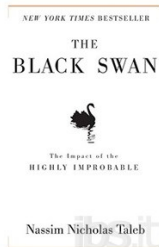
Laske keskihajonta.

Tapa 1 (määritelmästä):

$$\begin{aligned} \text{SD}(X) &= \sqrt{\sum_x (x - \mu)^2 f(x)} \\ &= \sqrt{(0 - 1)^2 \times (1 - 10^{-6}) + (10^6 - 1)^2 \times 10^{-6}} \approx 1000. \end{aligned}$$

Tapa 2 (laskentakaavan avulla):

$$\begin{aligned} \mathbb{E}(X^2) &= \sum_x x^2 f(x) = 0^2 \times (1 - 10^{-6}) + (10^6)^2 \times 10^{-6} = 10^6. \\ \implies \text{SD}(X) &= \sqrt{\mathbb{E}(X^2) - \mu^2} = \sqrt{10^6 - 1^2} \approx 1000. \end{aligned}$$



Esimerkki: Metro, jatkuva jakauma

Odotusaika X jatkuvasti tasajakautunut välillä $[0, 10]$. Sen odotusarvo $\mu = 5$ (minuuttia). Laske keskihajonta.

Tapa 1 (määritelmästä):

$$\text{SD}(X) = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx} = \sqrt{\int_0^{10} (x - 5)^2 \frac{1}{10} dx} = \dots$$

Tapa 2 (laskentakaavan avulla):

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{10} x^2 \frac{1}{10} dx = \frac{1}{10} \Big/_0^{10} \left(\frac{1}{3} x^3 \right) \approx 33.33.$$

$$\implies \text{SD}(X) = \sqrt{\mathbb{E}(X^2) - \mu^2} = \sqrt{33.33 - 5^2} \approx 2.89 \text{ minuuttia.}$$

Esimerkki: Huoneluku Suomen asunnoissa

(Livedemo)

Siirretyn ja skaalatun satunnaisluvun keskihajonta

Fakta (Viime luento)

- (i) $\mathbb{E}(a) = a$.
- (ii) $\mathbb{E}(bX) = b\mathbb{E}(X)$.
- (iii) $\mathbb{E}(X + a) = \mathbb{E}(X) + a$.

Fakta

- (i) $SD(a) = 0$.
- (ii) $SD(bX) = |b| SD(X)$.
- (iii) $SD(X + a) = SD(X)$.

Todistus.

(i) on helppo. Todistetaan (ii). Merkitään $\mu = \mathbb{E}(X)$.

$$\begin{aligned}\text{Var}(bX) &= \mathbb{E}[(bX - \mathbb{E}(bX))^2] = \mathbb{E}[(bX - b\mu)^2] \\ &= \mathbb{E}[b^2 (X - \mu)^2] = b^2 \mathbb{E}[(X - \mu)^2] = b^2 \text{Var}(X),\end{aligned}$$

joten

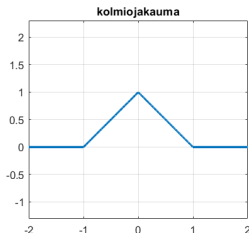
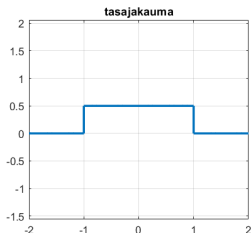
$$SD(bX) = \sqrt{\text{Var}(bX)} = \sqrt{b^2 \text{Var}(X)} = |b| SD(X).$$

(iii) samaan tapaan, kokeile itse!



Laskutehtävä: Tasa- ja kolmiojakauma

X on tasajakautunut välillä $[-1, 1]$, tiheys $f_X(x) = 0.5$ tällä välillä.
 Y on kolmiojakautunut samalla välillä, tiheys $f_Y(y) = 1 - |y|$ tällä välillä.



Poll: Arvaa onko keskihajonnoissa eroa.

Tehtävä: Laske molemmat keskihajonnat kaavalla

$$SD(X) = \sqrt{\mathbb{E}[(X - \mu_X)^2]}.$$

Huomaa että tässä on $\mu_X = \mu_Y = 0$. Tarvitset integrointia.

Sisältö

Varianssi ja keskihajonta

Poikkeaman todennäköisyyden yläraja

Kovarianssi ja korrelaatio

Tšebyšov in epäyhtälö: Poikkeamat odotusarvosta

Fakta (Tšebyšov in epäyhtälö, Chebyshev's inequality)

Jokaiselle satunnaisluvulle odotusarvona μ ja keskihajontana σ , tapahtuman

$$\{X = \mu \pm 2\sigma\} = \{\mu - 2\sigma \leq X \leq \mu + 2\sigma\}$$

todennäköisyys on vähintään

$$\mathbb{P}(X = \mu \pm 2\sigma) \geq \frac{3}{4}.$$



Pafnuti Tšebyšov
(engl. Pafnuty
Chebyshev)
1821–1894

Yleisemmin $\mathbb{P}(X = \mu \pm r\sigma) \geq 1 - \frac{1}{r^2}$ kaikilla $r \geq 1$.

- X :n arvo sijaitsee melko todennäköisesti (tn $\geq 75\%$) kahden keskihajonnan sisällä odotusarvostaan
- X :n arvo sijaitsee hyvin todennäköisesti (tn $\geq 99\%$) kymmenen keskihajonnan sisällä odotusarvostaan

Tšebyšov in epäyhtälö antaa keskiosan tn:lle alarajan (ja häntätodennäköisyydelle ylärajan). Jos jakauman muoto tiedetään, voidaan saada tiukempiakin rajoja.

Esimerkki: Dokumenttien pituudet

Eräässä lehdessä artikkelien sanamäärällä on keskiarvo 1000 ja keskihajonta 200. **Emme tunne jakaumaa sen tarkemmin.** Onko todennäköistä, että satunnaisen artikkelin sanamäärä on

- (a) välillä [600, 1400]? (2 keskihajonnan sisällä keskiarvosta)
- (b) välillä [800, 1200]? (1 keskihajonnan sisällä keskiarvosta)

Ratkaisu

- (a) Tšebyšov in epäyhtälöstä

$$\mathbb{P}(X \in [600, 1400]) = \mathbb{P}(X = \mu \pm 2\sigma) \geq 75\%,$$

joten sanamäärä on melko todennäköisesti välillä 600–1400 kB.

- (b) Tšebyšov ei tällä tarkkuudella kerro mitään hyödyllistä, sillä

$$\mathbb{P}(X \in [800, 1200]) = \mathbb{P}(X = \mu \pm \sigma) \geq 1 - \frac{1}{1^2} = 0.$$

Tarvitsisimme tarkempaa tiedon jakauman muodosta.

Esimerkki: Dokumenttien pituudet (jos normaalijakauma)

Eräässä lehdessä artikkelien sanamäärällä on keskiarvo 1000 ja keskihajonta 200. **Tiedämme lisäksi, että jakauma on ns. normaalijakauma.** Onko todennäköistä, että satunnaisen artikkelin sanamäärä on

- (a) välillä [600, 1400]? (2 keskihajonnan sisällä keskiarvosta)
- (b) välillä [800, 1200]? (1 keskihajonnan sisällä keskiarvosta)

Ratkaisu

- (a) Normaalijakauman taulukoista (tai R:llä $1-2*pnorm(-2)$)

$$\mathbb{P}(X \in [600, 1400]) = \mathbb{P}(X = \mu \pm 2\sigma) = \mathbb{P}\left(\frac{X - \mu}{\sigma} = 0 \pm 2\right) \approx 95\%.$$

- (b) Normaalijakauman taulukoista (tai R:llä $1-2*pnorm(-1)$)

$$\mathbb{P}(X \in [800, 1200]) = \mathbb{P}(X = \mu \pm \sigma) = \mathbb{P}\left(\frac{X - \mu}{\sigma} = 0 \pm 1\right) \approx 68\%.$$

Saimme paljon korkeampia rajoja, koska tiedämme jakauman.

Esimerkki: Dokumenttien pituudet (kolmas tapaus)

Eräissä lehdessä artikkelien sanamäärällä on keskiarvo 1000 ja keskihajonta 200. Onko todennäköistä, että satunnaisen artikkelin sanamäärä on

- (a) välillä [600, 1400]? (2 keskihajonnan sisällä keskiarvosta)
- (b) välillä [800, 1200]? (1 keskihajonnan sisällä keskiarvosta)

kun dokumenttien pituusjakauma on

k	750	1000	1250
$\mathbb{P}(X = k)$	32%	36%	32%

Ratkaisu

Suoraan jakauman taulukosta nähdään, että pituus on

- (a) varmasti (tn = 100%) välillä $\mu \pm 2\sigma = [600, 1400]$, mutta
- (b) melko epätodennäköisesti (tn = 36%) välillä $\mu \pm \sigma = [800, 1200]$.

Pohdittavaksi: Miten esimerkkiluvut valittiin? Haluttiin jakauma, jolla on $SD=200$, ja kaksi yhtä todennäköistä arvoa odotusarvon molemmin puolin. Miten valita niiden tn:t niin että SD on juuri se mitä haluttiin?

Tšebyšov in todistus (jatkuva; diskreetti samaan tapaan)

Valitaan mikä tahansa $r > 0$. Olkoon X :llä tiheys $f(x)$, odotusarvo μ ja keskihajonta σ . Olkoon MID väli $[\mu - r\sigma, \mu + r\sigma]$ ja TAIL sen komplementti. Nyt

$$\begin{aligned}\text{Var}(X) &= \sigma^2 = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx = \int_{\text{MID}} (\dots) + \int_{\text{TAIL}} (\dots) \\ &\geq \int_{\text{TAIL}} (x - \mu)^2 f(x) dx \geq \int_{\text{TAIL}} (r\sigma)^2 f(x) dx \\ &= r^2 \sigma^2 \int_{\text{TAIL}} f(x) dx = r^2 \sigma^2 \mathbb{P}(X \in \text{TAIL}).\end{aligned}$$

Kumotaan σ^2 ja siirretään r^2 toiselle puolelle:

$$\mathbb{P}(X \in \text{TAIL}) \leq \frac{1}{r^2}.$$

Huom. Tšebyšov in avulla voidaan todistaa suurten lukujen laki. Siihen tarvitaan vielä yksi väline, nimittäin satunnaismuuttujien summan varianssi; ks. seuraava luento ja https://en.wikipedia.org/wiki/Law_of_large_numbers

Sisältö

Varianssi ja keskihajonta

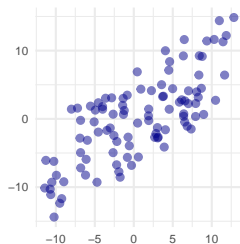
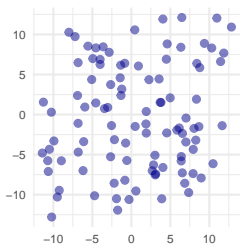
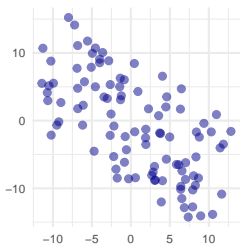
Poikkeaman todennäköisyyden yläraja

Kovarianssi ja korrelaatio

Yhteisvaihtelu

Keskihajonta mittaa yhden satunnaismuuttujan vaihtelua odotusarvonsa ympärillä.

Miten mitataan kahden satunnaismuuttujan X ja Y yhteisvaihtelua (suunta ja voimakkuus)?



Kovarianssi (engl. *covariance*)

$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$, mittaa satunnaismuuttujien X :n ja Y :n yhteisvaihtelun suuntaa ja voimakkuutta.

Diskreetti yhteisjakauma:

$$\sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y)$$

Jatkuva yhteisjakauma:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

Kovarianssi

- on > 0 , kun $X - \mu_X$ ja $Y - \mu_Y$ ovat usein samanmerkkiset
- on < 0 , kun $X - \mu_X$ ja $Y - \mu_Y$ ovat usein erimerkkiset
- yksiköltään alkup. muuttujien yksiköiden tulo (esim. m^2 , $\text{kg} \cdot \text{m}$, ...)

Kovarianssia ei normiteta ottamalla neliöjuurta (miksi)?

(Voi olla negatiivinen, eikä sen yksikkökään välttämättä ole neliö)

Huom. erikoistapaus:

$$\text{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)] = \mathbb{E}[(X - \mu_X)^2] = \text{Var}(X).$$

Kovarianssi: Vaihtoehtoinen laskentakaava

Tämä on usein laskuissa kätevämpi kuin määritelmän kaava.
Tämänkin voi laskea odotusarvon muunnoskaavalla.

Fakta

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Todistus.

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mathbb{E}[\mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_X \mu_Y.\end{aligned}$$



Kovarianssin bilineaarisuus

Fakta

Kovarianssioperaattori $(X, Y) \mapsto \text{Cov}(X, Y)$ on symmetrinen ja bilineaarinen (lineaarinen molempien argumenttiensa suhteen):

$$\text{Cov}(Y, X) = \text{Cov}(X, Y)$$

$$\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y).$$

$$\text{Cov}(X, Y_1 + Y_2) = \text{Cov}(X, Y_1) + \text{Cov}(X, Y_2).$$

$$\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$$

Yleisesti:

$$\text{Cov} \left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

Kovarianssin bilinearisuus: Todistus

Merkitään $Y = \sum_{j=1}^n b_j Y_j$. (i) Kovarianssin laskentakaavasta ja odotusarvon lineaarisuudesta

$$\begin{aligned}\text{Cov}\left(\sum_i a_i X_i, Y\right) &= \mathbb{E}\left[\left(\sum_i a_i X_i\right)Y\right] - \mathbb{E}\left[\left(\sum_i a_i X_i\right)\right]\mathbb{E}[Y] \\ &= \sum_i a_i \mathbb{E}[X_i Y] - \left(\sum_i a_i \mathbb{E}[X_i]\right) \mathbb{E}[Y] \\ &= \sum_i a_i \mathbb{E}[X_i Y] - \sum_i a_i \mathbb{E}[X_i] \mathbb{E}[Y] \\ &= \sum_i a_i (\mathbb{E}[X_i Y] - \mathbb{E}[X_i] \mathbb{E}[Y]) = \sum_i a_i \text{Cov}(X_i, Y).\end{aligned}$$

Symmetrian ja kohdan (i) avulla

$$\begin{aligned}\sum_i a_i \text{Cov}(X_i, Y) &= \sum_i a_i \text{Cov}(Y, X_i) \\ &= \sum_i a_i \text{Cov}\left(\sum_j b_j Y_j, X_i\right) \\ &= \sum_i a_i \sum_j b_j \text{Cov}(Y_j, X_i) \\ &= \sum_i \sum_j a_i b_j \text{Cov}(X_i, Y_j).\end{aligned}$$

Kovarianssi: Yhteenveto

Satunnaislukujen X ja Y kovarianssi on

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

missä $\mu_X = \mathbb{E}(X)$ ja $\mu_Y = \mathbb{E}(Y)$.

Diskreetti yhteisjakauma:

Jatkuva yhteisjakauma:

$$\sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy.$$

Kovarianssi on symmetrinen ja bilineaarinen:

$$\text{Cov}(Y, X) = \text{Cov}(X, Y)$$

$$\text{Cov} \left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

Korrelaatio (engl. *correlation*)

Kovarianssia ei normiteta ottamalla neliöjuurta (mm. koska se voi olla negatiivinen, ja yksikkökin voi olla esim. kg m)

Lisäksi haluaisimme luvun, joka kuvaa kovarianssia *suhteessa* siihen, paljonko X ja Y yleensäkin vaihtelevat. Siksi toisenlainen normitus . . .

Korrelaatio

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

mittaa satunnaislukujen X ja Y yhteisvaihtelun suuntaa ja voimakkuutta normitetuissa yksiköissä.

Voidaan osoittaa, että aina on $-1 \leq \text{Cor}(X, Y) \leq +1$.
(Todistus Cauchy-Schwarzin epäyhtälöllä, ei tällä kurssilla.)

Riippumattomat satunnaisluvut eivät korreloi

Fakta

Jos X ja Y ovat stokastisesti riippumattomat, niin

$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ ja $\text{Cor}(X, Y) = 0$.

Todistus.

Diskreetti.

$$\begin{aligned}\mathbb{E}(XY) &= \sum_x \sum_y xy f_{X,Y}(x, y) \\ &= \sum_x \sum_y xy f_X(x) f_Y(y) \\ &= \left(\sum_x x f_X(x) \right) \left(\sum_y y f_Y(y) \right) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Kovarianssin laskukaavasta

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

Siis myös $\text{Cor}(X, Y) = 0$. □

Esimerkki: Kaksi binaarista satunnaismuuttujaa

X ja Y ovat joukossa $\{-1, +1\}$ tasajakautuneita.

Lisäksi $c = \mathbb{P}(X = +1, Y = +1)$.

Määritä X :n ja Y :n yhteisjakauma ja korrelaatio.

	Y		
X	-1	+1	Yht
-1	c	$\frac{1}{2} - c$	$\frac{1}{2}$
+1	$\frac{1}{2} - c$	c	$\frac{1}{2}$
Yht	$\frac{1}{2}$	$\frac{1}{2}$	

$$\mathbb{E}(X) = 0$$

$$\mathbb{E}(X^2) = (-1)^2 \times \frac{1}{2} + (+1)^2 \times \frac{1}{2} = 1$$

$$SD(X) = \sqrt{\mathbb{E}(X^2) - (\mathbb{E}(X))^2} = \sqrt{1 - 0^2} = 1$$

$$\mathbb{E}(Y) = \mathbb{E}(X) = 0, \quad SD(Y) = SD(X) = 1.$$

$$\mathbb{E}(XY) = (-1)^2 \times c + 2 \times (-1)(+1) \times \left(\frac{1}{2} - c\right) + (+1)^2 c = 4c - 1$$

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 4c - 1$$

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)} = 4c - 1$$

Esimerkki: Asuntokunnat huone- ja henkilöluvun mukaan

(X =henkilöluku, Y =huoneluku)

		X						
		1	2	3	4	5	6	yht
Y	1	0.126	0.013	0.002	0.001	0.000	0.000	0.142
	2	0.196	0.086	0.012	0.005	0.001	0.000	0.301
	3	0.073	0.097	0.034	0.019	0.005	0.001	0.228
	4	0.038	0.079	0.031	0.030	0.010	0.003	0.191
	5	0.015	0.041	0.017	0.021	0.009	0.002	0.105
	6	0.004	0.012	0.006	0.007	0.003	0.001	0.032
yht		0.453	0.328	0.101	0.082	0.029	0.008	1.000

(Enemmän käsittelyä videolla, ks. luentovideo.)

Esimerkki: Lineaarinen deterministinen riippuvuus

Oletetaan, että erälle sm:ille X ja Y pätee **täsmälleen** $Y = a + bX$, missä X noudattaa jotain (tunnettua tai tuntematonta) jakaumaa odotusarvona $\mathbb{E}(X) = \mu$ ja keskihajontana $SD(X) = \sigma$. (Huom. deterministisyys.)

Laske tässä tilanteessa X :n ja Y :n korrelaatio.

$$\text{Cov}(X, Y) = \text{Cov}(X, a + bX) = \text{Cov}(X, a) + \text{Cov}(X, bX) = b\text{Var}(X).$$

$$SD(Y) = SD(a + bX) = |b| SD(X)$$

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)} = \frac{b\text{Var}(X)}{|b|SD(X)^2} = \frac{b}{|b|}.$$

$$\text{Cor}(X, Y) = \begin{cases} +1, & b > 0, \\ 0, & b = 0, \\ -1, & b < 0. \end{cases}$$

Varianssin yhteenlaskukaava

Jos X ja Y ovat satunnaismuuttujia (saavat olla riippuvia!), kovarianssin bilineaarisuudesta voidaan johtaa:

$$\begin{aligned}\text{Var}(X + Y) &= \text{Cov}(X + Y, X + Y) \\ &= \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y, Y) \\ &= \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y).\end{aligned}$$

Tämä on (yleinen) **varianssin yhteenlaskukaava**. Huomaa “ristitermi” eli “kovarianssitermi”. (Tällaista ei ollut odotusarvon yhteenlaskukaavassa.)

Varianssin yhteenlaskukaava riippumattomilla sm.

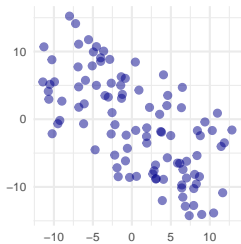
Jos X ja Y ovat stokastisesti **riippumattomat**, niin kovarianssitermi on nolla, eli siinä tapauksessa saadaan

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

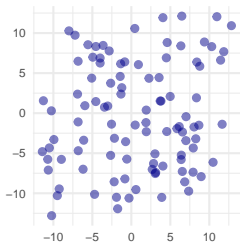
Yleisemmin, jos X_1, \dots, X_n ovat stokastisesti riippumattomat, pätee

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i).$$

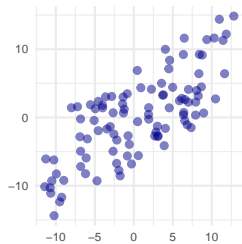
Yhteisjakaumasta simuloituja lukupareja (x, y)



$$\rho = -0.60$$



$$\rho = 0.28$$



$$\rho = 0.80$$

Seuraavalla kerralla puhutaan satunnaismuuttujien summista ja normaaliapproksimaatiosta. . .