5 Korrelaatio

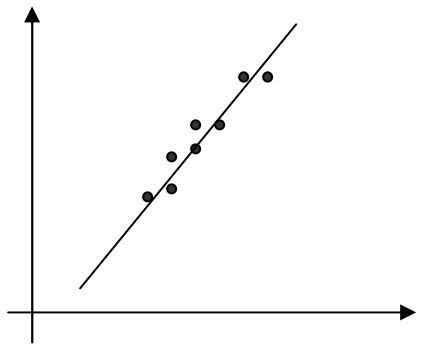
Jossakin piireissä saatetaan olla kiinnostuneita siitä, onko vanhempien runsaalla alkoholinkäytöllä yhteyttä lapsen koulumenestykseen. Tämäntapaisen ongelman tutkijalle löytyy helposti yksittäisiä esimerkkejä, sillä joku tuntee Jorman, jonka kotona ei käytetä alkoholia lainkaan, ja Jormakin menestyy koulussa varsin hyvin. Joku toinen taikka saattaa samakin henkilö tuntea Maurin ja tietää, että Maurin isä nappaa viinaa aika herkästi ja äitiä hoidetaan alkoholistiparantolassa. Silti myös Mauri menestyy koulussa erinomaisesti.

Pelkästään yhden ja kahden esimerkin nojalla ei voi eikä saa tehdä (vaikka torimuijat, joskus korkeastikin koulutetut herrat tekevät) varmoja päätelmiä vanhempien alkoholinkäytön ja lasten koulumenestyksen välisestä riippuvuudesta. Jos asiasta halutaan niin varmaa (tilasto)tietoa, että voidaan esimerkiksi ryhtyä yhteiskunnan taholta valistustyöhön tai voimatoimiin, on tutkittava suuri joukko koteja. Tällaisen tutkimuksen suorittaminen ei ole vallan pieni asia, sillä pelkkä aineiston keruu saattaa kestää vuosia.

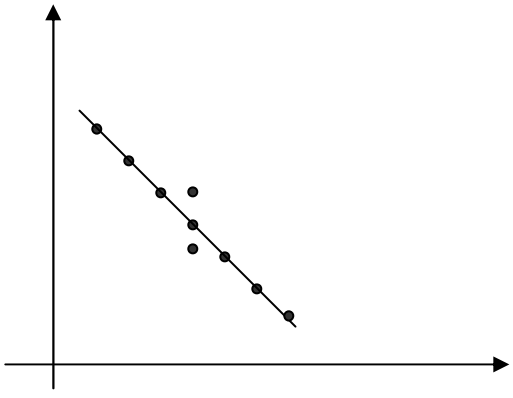
Mittaustuloksiin kohdistuneen mittavan matemaattisen käsittelyn jälkeen saatetaan löytää tarkkaakin tietoa siitä, onko riippuvuutta vai eikö sitä ole. Tällaisesta tilastollisesta riippuvuudesta käytetään nimitystä **korrelaatio**.

Seuraavan sivun kuviossa (kuva 5) on kuvattu uskonnossa ja historiassa menestymistä. Tilastoyksikköön a_i liittyvät havainnot x_i ja y_i on esitetty xy-koordinaatiston pisteinä. Jos tilastoyksiköllä nimeltä Minna on uskonnossa 9 ja historiassa 8, Minna esiintyy sitten koordinaatistossa pisteenä (9,8)

Kuvassa 5 pisteet ovat hajallaan erään nousevan suoran ympärillä. Korrelaation sanotaan olevan positiivista, kun x:n kasvaessa myös y kasvaa. Korrelaatio on tässä vielä melko korkeakin ts. suuri x merkitsee suurta y:tä; vieläpä toisen muuttujan arvo ennustaa melko hyvin toisenkin. Siis uskonnossa hyvä on myös historiassa hyvä.



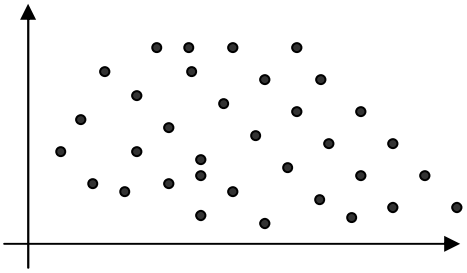
Kuva 5. Positiivinen korrelaatio



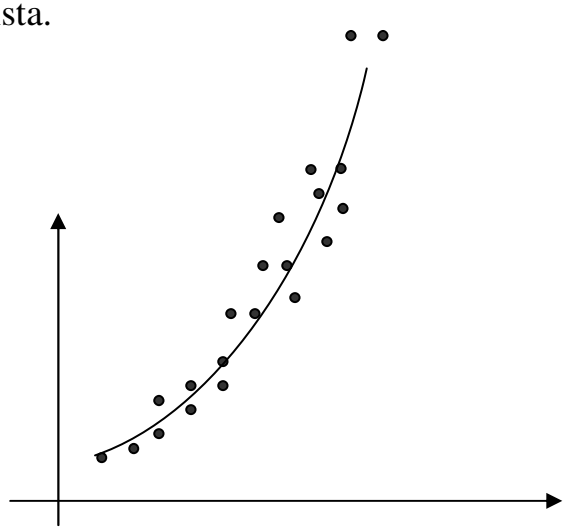
Kuva 6. Negatiivinen korrelaatio

Kuvassa 6 pistejoukko on sijoittunut laskevan suoran ympäristöön. Korrelaatio on nyt negatiivista ts. x:n kasvaessa y yleensä vähenee. Korrelaatio on tässä tapauksessa voimakkaampaa kuin kuvan 5 tilanteessa.

Kuvan 7 tapauksessa pistejoukko on niin hajaantunut, ettei toisen muuttujan arvo ennusta toista minkään vertaa. Sanotaan, ettei korrelaatiota ole tai että korrelaatio on nolla. Tällaista riippuvuutta voisi esiintyä vaikkapa koehenkilön massan ja asuinpaikan vuotuisen keskilämpötilan välillä. Kuvan 8 tapauksessa korrelaatio on varsin voimakasta, muttei se ole lineaarista.



Kuva 7. Ei korrelaatiota



Kuva 8. Eksponentiaalinen korrelaatio

Korrelaatiota voidaan tutkia paitsi graafisesti kuten edellä kuvista, myös laskennollisesti. Tällä kurssilla ei puututa kuitenkaan käyräviivaisiin korrelaatioihin eikä niiden matematiikkaan, vaan rajoitutaan lineaarisiin tapauksiin, joissa siis korrelaation esiintyessä pisteet sijaitsevat jonkun suoran tuntumassa. Määritellään tutkittavien ominaisuuksien välille **korrelaatiokerroin** r, jolla on sellaiset ominaisuudet, että

- r = 1, jos havaintopisteet sijaitsevat täydellisesti jollakin nousevalla suoralla,
- r = – 1, jos pisteet sijaitsevat täydellisesti jollakin laskevalla suoralla ja
- r = 0, jos pisteet ovat niin hajaantuneet, ettei voida havaita minkäänlaista riippuvuutta.

Näin korrelaatiokerroin rajataan välille $-1 \leq r \leq 1$ ja korrelaatio on sitä voimakkaampaa, mitä lähempänä ykköstä kertoimen itseisarvo on. Määritellään tällaisen korrelaatiokertoimen laskeminen:

****

MÄÄRITELMÄ 2:

Kuvatkoot x_i ja y_i tilastoyksikköön a_i liittyviä ominaisuuksia. Jos havainnot on **standardisoitu, siis** $z_{x_i} = \frac{x_i - \bar{x}}{\sigma_x}$ ja $z_{y_i} = \frac{y_i - \bar{y}}{\sigma_y}$, niin Pearsonin tulomomenttikerroin

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

taikka standardoimattomille pisteille

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot \sigma_x \cdot \sigma_y}$$

****

Korrelaatiokertoimen laskeminen on varsin työläs toimenpide, jos vallankin havaintoarvopareja on vähänkin suurempi määrä. Toki laskimia voidaan käytännön tilastoanalyyseissä hyödyntää, mutta korrelaatiokertoimen määritys on syytä hahmottaa, jotta laskiessaan tietää, mistä on kysymys. Korrelaatiokertoimen mekaanisen määrittämisen aivan niin kuin keskihajonnankin määrittämisen voi opetella suorittamaan laskimella ymmärtämättä vähääkään, mistä on kysymys.

Esim. 1 Kymmeneltä lukion päättäneeltä kysyttiin heidän päivittäinen keskimääräinen kotitehtävien suoritusaikansa (x) tunteina ja ylioppilastodistuksen puoltoäänien lukumäärä (y).
(3,22) (2,15) (3,28) (4,33) (1,25) (4,26) (5,27) (3,35) (2,19) (4,30).

Laskettava korrelaatiokerroin.

Kertoimen laskemiseksi tarvitaan kummankin muuttujan keskiarvo ja keskihajonta. Laaditaan sellainen taulukko, että nämä voidaan laskea.

Taulukko 5. Korrelaatiokertoimen laskeminen

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
3	22	0	0	-4	16	0
1	15	-2	4	-11	121	22
3	28	0	0	2	4	0
4	33	1	1	7	49	7
1	25	-2	4	-1	1	2
4	26	1	1	0	0	0
5	27	2	4	1	1	2
3	35	0	0	9	81	0
2	19	-1	1	-7	49	7
4	30	1	1	4	16	4
30	260		16		338	44
1,269	5.81					

$$\sigma_x = \sqrt{\frac{16}{10}} = 1.2649... \qquad \sigma y = \sqrt{\frac{338}{10}} = 5.81377...$$

$$r = \frac{44}{(10-1)\cdot 1.2649 \cdot 5.81377} = 0.6648.... \approx 0.66$$

Korrelaatiokertoimen tulkinta on sellainen, että korrelaatio on

- voimakas, jos $|r| \geq 0.8$
- huomattava, jos $0.6 \leq |r| < 0.8$
- kohtalainen, jos $0.3 \leq |r| < 0.6$
- merkityksetön, jos $|r| < 0.3$

Lienee ymmärrettävissä, ettei esimerkin tapauksessa korrelaatio ole täydellinen. Lahjakkaan henkilön ei tarvitse lukea mahdottomia päivittäisiä tuntimääriä hyviin, jopa erinomaisiin tuloksiin päästäkseen ja toisaalta heikolla oppilaalla kovatkaan ponnistelut eivät vie korkeisiin saavutuksiin.