

ESITIEDOT:

KATSO MYÖS: ■ todennäköisyyslaskenta, ■ todennäköisyysjakaumat, ■ keskiarvo

---

## Tilastodata

Tilastojen pohjana on jollakin johdonmukaisella tavalla kerätty jotakin ilmiötä koskeva *data* (tiedot). Esimerkkejä ovat tuhannelta hengeltä kysytty poliittinen puoluekanta, koululuokan matematiikan kokeen arvosanat ja joukko-osaston alokkaiden pituudet.

Data voi olla *diskreettiä*, jolloin vain tietyt arvot tulevat kysymykseen (puoluekanta, arvosanat), tai *jatkuvaa*, jolloin periaatteessa mikä tahansa jollakin välillä oleva reaaliarvo on mahdollinen (ihmisten pituudet). Data voi olla *numeerista* (arvosanat, pituudet) tai *nominaalista* (puolueiden nimet). Jälkimmäisessä tapauksessa se voidaan koodata numeeriseen muotoon, mutta koodaus ei (välttämättä) merkitse järjestyksen muodostamista: vaikka puolueille annettaisiinkin numerot, nämä tuskin kuvaavat puolueiden sijoittumista vaikkapa oikeisto-vasemmisto -akselille.

Diskreetin datan tapauksessa voidaan jokaiselle data-arvolle laskea sen *frekvenssi*: montako kertaa arvo esiintyy datassa. Jatkuva data voidaan ensin luokitella (esimerkiksi sijoittaa ihmisten pituuksien arvot viiden senttimetrin pituisille väleille) ja tämän jälkeen muodostaa luokkien frekvenssit.

Frekvenssit voidaan esittää havainnollisessa muodossa erilaisina diagrammeina. Tavallisimpia ovat *pylväsdiagrammit* eli *histogrammit*, joissa esitetään yleensä absoluuttiset frekvenssit, ja *piirakkadiagrammit*, joista ilmenevät frekvenssien suhteelliset osuudet kokonaisuudesta.

Jos data esittää ilmiön kehitystä ajan mukana, ts. kyseessä on *aikasarja*, voidaan ilmiötä kuvata ajan funktiona. Argumenttina vaaka-akselilla on tällöin aika.

■ funktio

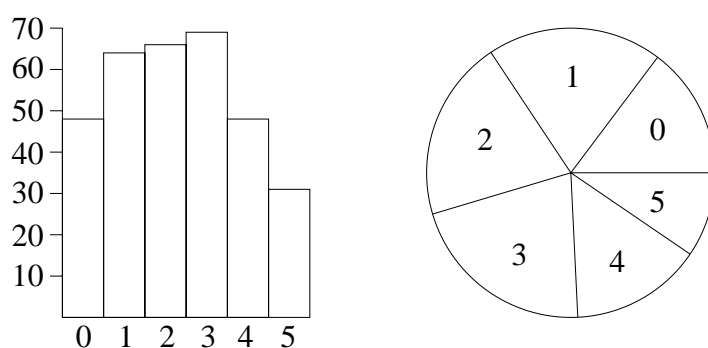
Esimerkkejä seuraavassa.

ESITIEDOT:

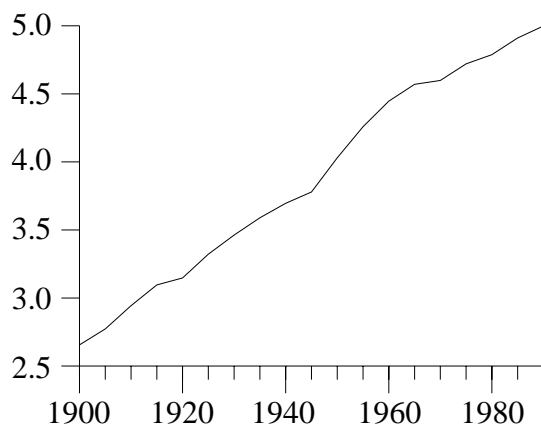
KATSO MYÖS: ■ todennäköisyyslaskenta, ■ todennäköisyysjakaumat, ■ keskiarvo

## Tilastodatan esittäminen

Esimerkkinä pylväs- ja piirakkadiagrammista olkoon Teknillisen korkeakoulun erään matematiikan peruskurssin arvosanajakauma. Pylväsdiagrammissa on vaaka-akselilla eri arvosanat (0 = hylätty, huonoin hyväksytty 1, paras 5); pystyakselilla ovat eri arvosanan saaneiden opiskelijoiden absoluuttiset määrät. Piirakkadiagrammista ilmenevät eri arvosanojen suhteelliset osuudet havainnollisesti.



Suomen väkilukudata muodostaa aikasarjan. Alla on väkiluvun kehitystä 1900-luvulla kuvaava funktio. Pystyakselilla on väkiluku miljoonina.



## Datan tunnusluvut

Datasta voidaan myös muodostaa *tunnuslukuja*, jotka kuvaavat joitakin sen oleellisia piirteitä. Tavallisimmat ovat seuraavat:

- *Moodi* on se datassa esiintyvä arvo, jota on lukumääräisesti eniten. Käyttö tulee kysymykseen vain diskreetin datan tapauksessa. Esimerkiksi kannatukseltaan suurin puolue tai yleisin koearvosana.
- *Mediaani* on suuruusjärjestykseen asetetun datan keskimäinen arvo (tai jos arvoja on parillinen määrä, kahden keskimäisen keskiarvo). Tällöin datan tulee olla numeerista.
- *Keskiarvo* voidaan laskea numeerisesta datasta. Jos data muodostuu luvuista  $x_1, x_2, \dots, x_n$ , näiden keskiarvo on

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k.$$

■ keskiarvo  
(aritmeettinen)  
■ summamer-  
kintä

- *Keskihajonta* kuvaa datan keskimääräistä poikkeamaa keskiarvosta. Todennäköisyyslaskennan teoriaan liittyvistä syistä lasketaan poikkeamien neliöiden keskiarvon neliöjuuri:

■ keskihajonta  
■ todennäköi-  
syyslaskenta

$$s = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2} = \sqrt{\frac{1}{n} \left[ \sum_{k=1}^n x_k^2 - \frac{1}{n} \left( \sum_{k=1}^n x_k \right)^2 \right]}.$$

Jälkimmäinen lauseke on toisinaan edellistä kätevämpi, koska tällöin sekä keskiarvon että keskihajonnan laskemista varten tarvitaan ainoastaan datassa olevien lukujen summa ja neliöiden summa.

Keskiarvon ja keskihajonnan lausekkeet voidaan esittää myös frekvenssien avulla. Nämä johdetaan melko helposti edellä esitetyistä.

## Matemaattinen tilastotiede

Tilastodatan perusteella pyritään usein tekemään johtopäätöksiä laajemmasta joukosta, kuin mistä data on kerätty. Esimerkiksi puolueiden kannatusta äänioikeutetun väestön keskuudessa arvioidaan haastattelemalla tuhat sopivasti valittua henkilöä; tietyn ikäisen miespuolisen väestön pituusjakaumaa tutkitaan mittaamalla osa asevelvollisten kutsuntoihin osallistuvista; hehkulamppujen käyttöikäjakaumaa arvioidaan polttamalla loppuun varsin rajoitettu osa tuotannosta.

Tyypillistä tällaisissa tilanteissa on, että koko joukkoa — tarkasteltavaa populaatiota — ei voida tutkia joko tutkimuksen hankaluuden tai kalleuden tähden, sen takia että tutkimus tuhoaa populaation (hehkulamput), siksi että halutaan selvittää jokin yleinen ominaisuus rajallisella määrällä toistettavia kokeita tai jostakin muusta tällaisesta syystä.

Tällöin muodostetaan koko populaatiosta *otos*, so. tutkitaan vain rajallista määrää populaation yksilöitä ja pyritään täten saadun datan avulla tekemään yleisiä johtopäätöksiä. Otos valitaan jollakin satunnaismenettelyllä siten, että systemaattisia virheitä ei synny. Käytössä on useita erilaisia menetelmiä.

Koko populaation tutkimisessa otoksen avulla on pohjana *matemaattinen tilastotiede*, joka perustuu todennäköisyyslaskentaan.

■ todennäköisyyslaskenta

## Estimointi

Matemaattisen tilastotieteen perusprobleema on tutkittavaan ilmiöön liittyvän todennäköisyysjakauman tunnuslukujen määrittäminen.

Olkoon tarkasteltavana ilmiö, jonka oletetaan luonteensa perusteella noudattavan jotakin todennäköisyysjakaumaa, mutta jotkin jakauman parametrit — esimerkiksi tiheysfunktion lausekkeessa olevat vakiot — ovat tuntemattomia. Näiden määrittämiseksi voidaan muodostaa otos, jonka avulla pyritään arvioimaan eli *estimoimaan* parametreille sopivat arvot.

■ tiheysfunktio

Esimerkiksi voidaan olettaa, että ihmisten pituudet noudattavat normaalijakaumaa, mutta tämän odotusarvo  $\mu$  ja keskihajonta  $\sigma$  ovat tuntemattomia. Mittaamalla sopivasti valittu ihmisjoukko (otos), voidaan nämä estimoida.

■ jakauma  
(normaali-)

■ odotusarvo

■ keskihajonta

■ todennäköi-  
syyslaskenta■ summamer-  
kintä

Jos otosdataa merkitään  $x_1, x_2, \dots, x_n$ , saadaan todennäköisyyslaskennan teorian perusteella odotusarvon estimaatiksi (jakaumasta riippumatta) *otoskeskiarvo*

$$\hat{x} = \frac{1}{n} \sum_{k=1}^n x_k \quad (= \bar{x}).$$

Keskihajonnan estimaatti on vastaavasti *otoskeskihajonta*

$$\hat{s} = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}.$$

Tämä poikkeaa koko datan keskihajonnasta  $s$  siten, että summamerkin edessä on jakajana  $n - 1$  havaintojen lukumäärän  $n$  sijasta.

Tuntemattomat parametrit voivat olla muunkinlaisia. Niiden estimaateille johdetaan vastaavantyyppiset lausekkeet todennäköisyyslaskennan teorian avulla.

## Tilastollinen testaus

Jotakin ilmiötä tutkittaessa siitä tehdään usein *hypoteeseja* ja tilastodatan (otoksen) perusteella pyritään päättämään, pitävätkö hypoteesit paikkansa. Tätä sanotaan hypoteesien *tilastolliseksi testaamiseksi*.

Tällöin joudutaan varautumaan kahden eri lajin virheeseen: hypoteesi hyväksytään, vaikka se on väärä; hypoteesi hylätään, vaikka se on oikea.

Todennäköisyyslaskennan teorian avulla suunnitellaan testi (otoksen koko, sopivan testisuureen laskeminen), jolla hypoteesin voimassaolo voidaan tarkistaa siten, että kummankin lajin virheen todennäköisyys saadaan riittävän pieneksi.

■ todennäköisyyslaskenta

Yksikertainen esimerkki on arpanopan virheettömyyden testaaminen. Hypoteesina on tällöin, että eri silmäluvut ovat yhtä todennäköisiä. Noppaa heitetään useita kertoja, jolloin saadaan joukko heittotuloksia. Nämä muodostavat otoksen kaikista mahdollisista nopalla tehtävistä heitoista. Tilastomatematiikka on ongelmana on määrittää riittävä heittojen määrä ja muodostaa heittotuloksista testisuure, jonka arvon perusteella voidaan päättää, onko hypoteesia pidettävä oikeana esimerkiksi 99 prosentin varmuudella.

Toisena esimerkkinä olkoon tuote-erän kelvollisuuden testaaminen: Tietty osa tuotteista tutkitaan (mikä voi merkitä niiden rikkomista) ja tulosten perusteella lasketaan arvo sopivasti muodostetulle testisuurelle. Tämän perusteella päätetään, päästetäänkö erä markkinoille vai ei. Todennäköisyys, että virheellinen erä hyväksytään, on saatava riittävän pieneksi. Toisaalta virheettömän erän hylkäystodennäköisyys ei myöskään saa olla liian suuri.

ESITIEDOT:

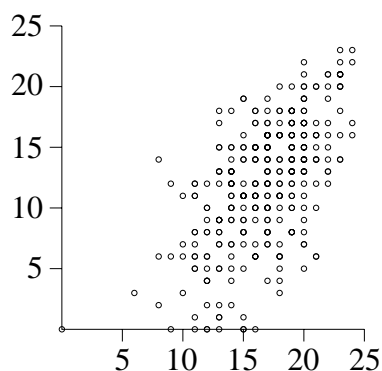
KATSO MYÖS: ■ todennäköisyyslaskenta, ■ todennäköisyysjakaumat, ■ keskiarvo

## Korrelaatio

Tutkimuksen kohteena olevasta populaatiosta tai siitä muodostetusta otoksesta voidaan koota kahta eri ominaisuutta koskeva data. Tällaisia voivat olla esimerkiksi jonkin koulun oppilaiden päästötodistusten matematiikan ja A-kielen arvosanat. Halutaan selvittää, millainen riippuvuus näiden välillä vallitsee: menestyvätkö matemaattisesti lahjakkaat yleensä hyvin myös kielissä vai keskittyvätkö oppilaat jompaankumpaan toisen jäädessä vähemmälle huomiolle.

Jos tarkastelussa on  $n$  oppilasta, joiden matematiikan arvosanat ovat  $x_1, x_2, \dots, x_n$  ja kieliarvosanat vastaavasti  $y_1, y_2, \dots, y_n$ , voidaan muodostaa graafinen esitys, jossa pisteet  $(x_k, y_k)$ ,  $k = 1, 2, \dots, n$ , sijoitetaan xy-koordinaatistoon. Tämä antaa jo jonkinlaisen kuvan tilanteesta.

Alla olevassa kuvassa on Teknillisen korkeakoulun matematiikan peruskurssin ensimmäisen ja toisen välikokeen tuloksista muodostettu graafinen esitys. Ensimmäisen välikokeen tulokset ovat vaaka-akselilla ja toisen pystyakselilla.



Näyttää ilmeiseltä, että välikoemenestysten välillä on positiivinen riippuvuus, ts. pisteet sijoittuvat nousevan suoran  $y = kx + b$ ,  $k > 0$ , ympärille.

■ kulmakerroin

Tarkempi mitta riippuvuudelle on korrelaatiokerroin; ks. seuraavaa.

## Korrelaatiokerroin

*Korrelaatiokerrointa* laskettaessa kumpikin data ensin skaalataan sen keskiarvolle ja keskihajonnalla, minkä jälkeen muodostetaan parittainen tulosumma:

■ summamer-  
kintä

$$r = \frac{1}{n} \sum_{k=1}^n \frac{x_k - \bar{x}}{s_x} \frac{y_k - \bar{y}}{s_y}.$$

Tässä  $\bar{x}$  ja  $\bar{y}$  tarkoittavat kummastakin datasta laskettuja keskiarvoja,  $s_x$  ja  $s_y$  ovat vastaavasti keskihajonnat. Lauseke voidaan saattaa myös muotoon

$$r = \frac{n \sum_{k=1}^n x_k y_k - (\sum_{k=1}^n x_k) (\sum_{k=1}^n y_k)}{\sqrt{n \sum_{k=1}^n x_k^2 - (\sum_{k=1}^n x_k)^2} \sqrt{n \sum_{k=1}^n y_k^2 - (\sum_{k=1}^n y_k)^2}},$$

jolloin tarvitsee laskea datasta vain summat, neliösummat ja tulosumma.

Korrelaatiokerroin on aina välillä  $[-1, 1]$ . Sen merkin mukaan puhutaan *positiivisesta* tai *negatiivisesta korrelaatiosta*. Mitä lähempänä ykköstä arvo on, sitä vahvempaa on ominaisuuksien esiintyminen yhdessä. Lähellä arvoa  $-1$  oleva korrelaatiokerroin osoittaa vastaavasti, että ominaisuudet eivät yleensä esiinny samanaikaisesti.

Vahvan positiivisen korrelaation sanotaan usein tarkoittavan ominaisuuksien välistä riippuvuutta. Tästä ei kuitenkaan voida päätellä, että ominaisuuksien välillä olisi kausaalisuhde, ts. toinen olisi toisen syy. Kyse on vain siitä, että ominaisuudet käyttäytyvät samansuuntaisesti. Niillä saattaa esimerkiksi olla jokin kolmas ominaisuus yhteisenä syynä.

Kun korrelaatiokerroin lasketaan otoksesta, joudutaan erikseen miettimään sen merkitsevyyttä. Otoksen epäedustavuus tai pienuus saattaa johtaa siihen, että esimerkiksi korrelaatiokertoimen arvon  $|r| \leq 0.5$  ei voida katsoa merkitsevästi poikkeavan nolasta.

Edellä olevassa Teknillisen korkeakoulun matematiikan peruskurssin välikokeita koskevassa esimerkissä välikoemenestysten välinen korrelaatiokerroin on 0.61.