Tilastollisia tunnuslukuja

Keskiarvo on eniten käytetty jakaumaa kuvaava tunnusluku. Muita usein käytettyjä jakaumaa kuvaavia tunnuslukuja ovat mediaani, moodi ja keskihajonta.

Esimerkki 1

Määritä lukujen 19, 10, 16, 12, 16, 13, 12, 16, 14 ja 11 mediaani, moodi ja keskiarvo.

Ratkaisu

Mediaani on suuruusjärjestykseen laitetun havaintoaineiston keskimmäinen arvo. Järjestetään luvut suuruusjärjestykseen.

10, 11, 12, 12, 13, 14, 16, 16, 16, 19

Keskimmäiset luvut ovat 13 ja 14. Mediaani on niiden keskiarvo $\frac{13+14}{2} = 13{,}5$.

Moodi on havaintoaineiston yleisin arvo. Aineistossa on

kolme kappaletta lukua 16 ja muita lukuja yksi tai kaksi kutakin. Moodi on siis luku 16.

Keskiarvo on
$\overline{x} = \frac{10 + 11 + 12 + 12 + 13 + 14 + 16 + 16 + 16 + 19}{10} = 13{,}9.$

Vastaus

Mediaani on 13,5, moodi 16 ja keskiarvo 13,9.

Esimerkki 2

Koululaisryhmä tutki koulun ohi ajavien autojen määrää eri viikonpäivinä joka aamu kello 8–9. Tulokset on koottu oheiseen taulukkoon. Määritä autojen lukumäärän keskiarvo ja keskihajonta. Poikkesiko jonakin päivänä autojen määrä keskiarvosta yli kahden keskihajonnan verran?

ma	ti	ke	to	pe
119	239	241	199	265

Avaa OpenOffice-tiedosto
→

Ratkaisu

Syötetään aineisto laskentaohjelmaan. Keskiarvoksi saadaan
$\overline{x} = 212{,}6$ ja keskihajonnaksi $\sigma = 51{,}371\ldots \approx 51{,}4$.

Lasketaan autojen lukumäärät, jotka ovat kahden keskihajonnan päässä keskiarvosta.

$212{,}6 + 2 \cdot 51{,}4 = 315{,}4$
$212{,}6 - 2 \cdot 51{,}4 = 109{,}8$

Koska autojen lukumäärä vaihteli välillä 119–265, autojen määrä ei poikennut yli kahta keskihajontaa keskiarvosta minään päivänä.

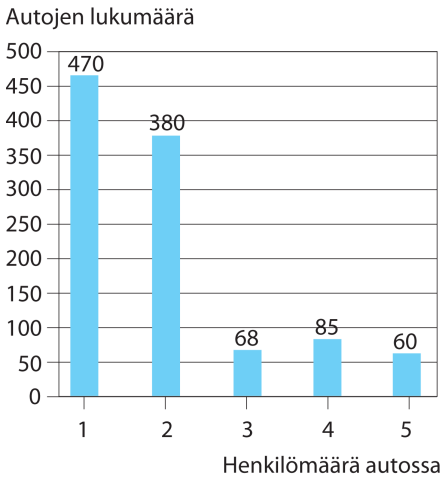
Vastaus

Keskiarvo on 212,6 ja keskihajonta 51,4. Lukumäärä ei poikennut keskiarvosta yli kahta keskihajontaa minään

päivänä.

Esimerkki 3

Koululaisryhmä tarkkaili liikennevaloissa henkilöautoja. Pylväskuvioon on koottu autojen määrät matkustajamäärän mukaan eroteltuna.



Laske suhteelliset frekvenssit henkilöiden lukumäärille ja laadi aineistosta ympyräkuvio.

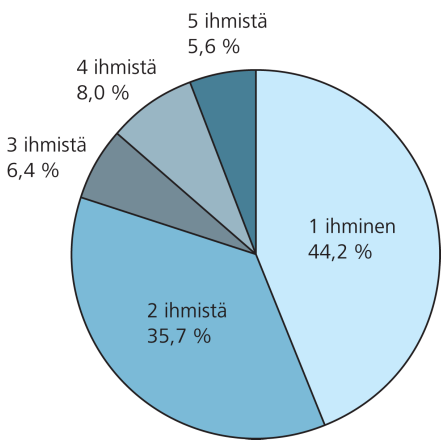
Ratkaisu

Syötetään aineisto taulukkolaskentaohjelmaan ja lasketaan autojen kokonaismäärä. Autoja on yhteensä 1 063.

Määritetään ohjelmalla myös suhteelliset frekvenssit.

Henkilömäärä autossa	***f***	***f* %**
1 ihminen	470	44,2 %
2 ihmistä	380	35,7 %
3 ihmistä	68	6,4 %
4 ihmistä	85	8,0 %
5 ihmistä	60	5,6 %

Piirretään ohjelmalla ympyräkuvio.



Korrelaatio

Korrelaatiokerroin kuvaa kahden tilastomuuttujan välistä lineaarista riippuvuutta. Korrelaatiokertoimen arvo vaihtelee välillä –1…1. Korrelaatiokerroin 1 tarkoittaa täydellistä positiivista lineaarista riippuvuutta ja –1 täydellistä negatiivista lineaarista riippuvuutta.

Esimerkki 4

Eteläinen Afrikka on maantieteellinen alue, johon kuuluu 12 valtiota Afrikan mantereella tai Intian valtameressä. Alueella on poikkeuksellisen alhainen eliniän ennuste. Alla olevaan taulukkoon on koottu tämän alueen hedelmällisyysluku, odotettavissa oleva elinikä ja lukutaito. Tutki alueen valtioiden
a) odotetun eliniän ja lukutaidon
b) hedelmällisyysluvun ja lukutaidon
välistä riippuvuutta ja tulkitse saamiasi tuloksia.

Maa	**Hedelmällisyys-luku (lasta)**	**Odotettu elinikä (v)**	**Lukutaito (%)**
Angola	5,4	54	71,2
Botswana	2,4	54	88,2
Etelä-Afrikka	2,3	56	94,6
Lesotho	2,8	50	79,4
Madagaskar	4,2	69	64,7
Malawi	5,0	57	66,0
Mauritius	1,5	74	90,6
Mosambik	4,9	53	58,8
Namibia	2,8	62	90,8
Sambia	5,5	51	85,1
Swazimaa	3,1	49	87,5

Zimbabwe	3,2	57	86,9

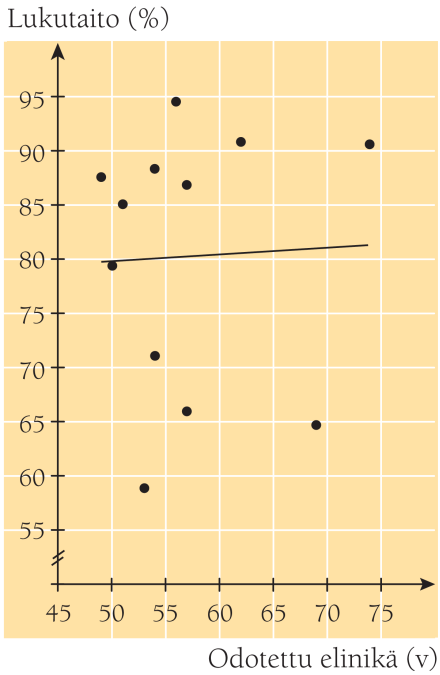
Avaa OpenOffice-tiedosto →

- Hedelmällisyysluku ilmoittaa, kuinka monta lasta nainen keskimäärin synnyttää.
- Odotettu elinikä kuvaa vastasyntyneen lapsen elinikää.
- Lukutaito ilmoittaa, kuinka monta prosenttia 15-vuotta täyttäneistä henkilöistä osaa lukea ja kirjoittaa.

Ratkaisu

Syötetään aineisto laskentaohjelmaan.

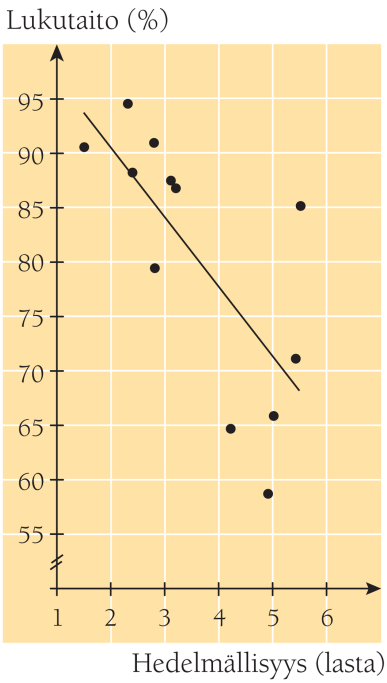
a) Sovitetaan odotetun eliniän ja lukutaidon välistä riippuvuutta kuvaava regressiosuora hajontakuvioon.



Määritetään lineaarisen riippuvuuden suuruutta kuvaava korrelaatiokerroin. Odotetun eliniän ja lukutaidon väliseksi korrelaatiokertoimeksi saadaan $r = 0{,}03515\ldots \approx 0{,}035$. Tämä tarkoittaa, että tällä alueella lukutaidon ja odotettavissa olevan eliniän välillä ei ole lineaarista riippuvuutta.

b) Sovitetaan hedelmällisyysluvun ja lukutaidon välistä riippuvuutta kuvaava regressiosuora hajontakuvioon.

Määritetään lineaarisen riippuvuuden suuruutta kuvaava korrelaatiokerroin. Hedelmällisyysluvun ja lukutaidon väliseksi korrelaatiokertoimeksi saadaan $r = -0{,}71614\ldots \approx -0{,}72$. Tämä tarkoittaa, että tällä alueella hedelmällisyysluvun ja lukutaidon välillä on huomattava negatiivinen korrelaatio. Maissa, joissa hedelmällisyysluku on korkea, lukutaitoisten osuus on pienempi.



Vastaus

a) Odotetun eliniän ja lukutaidon välillä ei ole tilastollista riippuvuutta. Korrelaatiokerroin on $r = 0{,}035$.
b) Hedelmällisyysluvun ja lukutaidon välillä on huomattava negatiivinen korrelaatio. Korrelaatiokerroin on $r = -0{,}72$.

Regressiokäyriä

Tilastollinen riippuvuus voi olla muutakin kuin lineaarista. Riippuvuus voi olla esimerkiksi eksponentiaalinen, toisen asteen polynominen tai jonkun muun asteen polynominen.

Esimerkki 5

Taulukossa on esitetty 100 metrin juoksun maailmanennätykset eri ikäluokissa.

Ikä (v)	Aika (s)	Ikä (v)	Aika (s)	Ikä (v)	Aika (s)
5	16,19	**10**	12,08	**15**	10,20
6	14,30	**11**	11,95	**16**	10,15
7	13,46	**12**	11,22	**17**	10,01

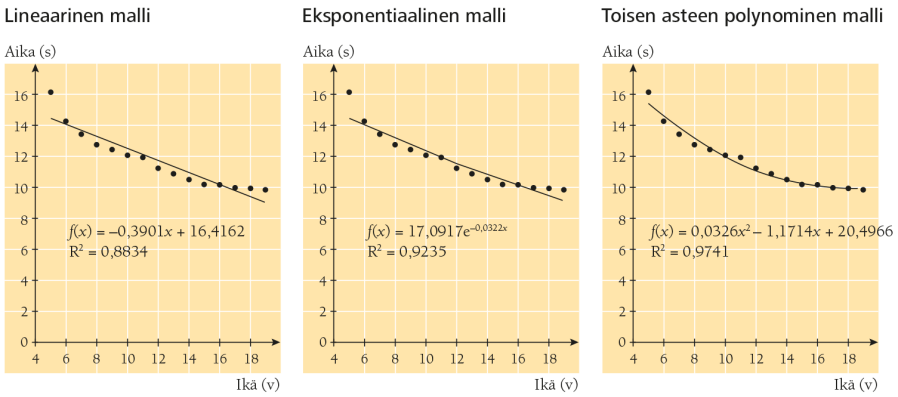
8	12,80	**13**	10,90	**18**	9,97
9	12,45	**14**	10,51	**19**	9,84

Avaa OpenOffice-tiedosto →

a) Sovita aineistoon lineaarinen, eksponentiaalinen ja toisen asteen polynominen malli.
b) Mikä malleista kuvaa aineistoa parhaiten välillä 5–19 vuotta?

Ratkaisu

a) Syötetään aineisto laskentaohjelmaan. Sovitetaan havaintoihin
lineaarinen, eksponentiaalinen ja toisen asteen polynominen
malli. Määritetään asetuksissa, että ohjelma antaa riippuvuutta
kuvaavan yhtälön ja sen selitysasteen R^2.



b) Korkein selitysaste $R^2 = 0{,}9741$ on toisen asteen polynomisella mallilla. Siinä havaintopisteet sijoittuvat parhaiten regressiokäyrälle.

TEORIAYHTEENVETO

Tilastollisia tunnuslukuja

- Keskiarvo on havaintoarvojen summa jaettuna niiden lukumäärällä.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Mediaani on järjestetyistä havaintoarvoista keskimmäinen. Jos havaintoarvot ovat lukuja ja niitä on parillinen määrä, mediaani on kahden keskimmäisen havaintoarvon keskiarvo.

- Moodi eli tyyppiarvo on yleisin muuttujan arvo.

- Havaintoarvojen $x_1, x_2, \ldots, x_n$ keskihajonta on

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}},$$

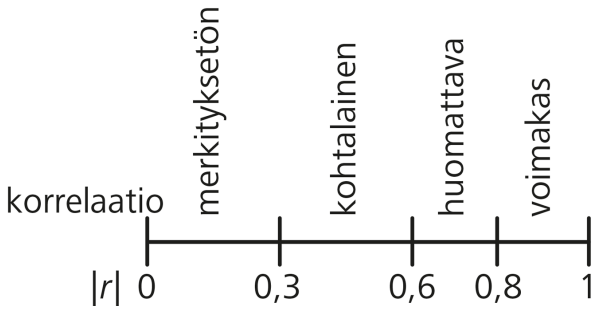
missä $\overline{x}$ on havaintoarvojen keskiarvo ja n havaintoarvojen lukumäärä.

Avaa appletti →

Tilastollinen riippuvuus

- Kahden tilastomuuttujan välisen lineaarisen riippuvuuden suuruutta mitataan korrelaatiokertoimella.

- Korrelaation voimakkuus määräytyy kertoimen itseisarvon perusteella, joten kertoimen etumerkki ei vaikuta korrelaation voimakkuuteen.



- Tilastomuuttujien välillä voi olla muukin kuin lineaarinen riippuvuus. Se voi olla esimerkiksi eksponentiaalinen tai toisen asteen polynominen riippuvuus.

- Selitysaste R^2 ilmoittaa, kuinka hyvin malli kuvaa aineistoa.

Avaa appletti →

LASKIMET JA LASKENTAOHJELMAT

• Kokeen A-osan peruslaskimella voidaan määrittää tilastollisia tunnuslukuja. Aineisto syötetään näppäimen DAT avulla painamalla näppäintä kunkin syötetyn havaintoarvon jälkeen. Keskiarvon saa näppäimellä MEA, keskihajonnan näppäimellä σ_N ja mediaanin näppäimellä MED.

• Kokeen B-osan laskentaohjelmat laskevat yleensä kaikki tilastolliset

tunnusluvut samalla kertaa. Havaintoarvot syötetään tai ladataan ensin ohjelman taulukkonäkymään. Alue maalataan ja käytetään tunnuslukujen määritystoimintoa, joka voi olla esimerkiksi *Yhden muuttujan analyysi*, *Yhden muuttujan tilastot* tai *Tilastotiedot/Tunnusluvut*. Geometriaohjelmaa käytettäessä pitää lopuksi vielä painaa näppäintä $\sum x$.

• Kaikkien ohjelmien tilastotoiminnot eivät määritä moodia. Taulukkolaskentaohjelma ei osaa tulkita aineistoa, jossa on frekvenssejä, joten tällaisen aineiston kohdalla kannattaa käyttää jotain muuta ohjelmaa.

• Tilastokuvioiden piirtämiseen kannattaa yleensä käyttää taulukkolaskentaohjelmaa, jossa toiminto on *Lisää/Kaavio*. Pylväsdiagrammeja ja kahden muuttujan hajontakuvioita voi piirtää myös geometriaohjelmalla.

• Korrelaatiokertoimen saa kahden muuttujan tilastotoiminnolla, joka voi olla esimerkiksi *Kahden muuttujan regressioanalyysi* tai *Lineaarinen regressio*. Taulukkolaskentaohjelmassa korrelaatiokertoimelle on oma funktionsa.

• Mallien vertailussa käytettävä selitysaste R^2 on yleensä mukana ohjelman laskemissa tunnusluvuissa. Taulukkolaskentaohjelmassa selitysasteen saa valittua näkyviin, kun sovittaa hajontakuvioon trendiviivan.