

Tilastoaineiston käsittely

Esimerkki 1

Havaintomatriisiin on koottu tiedot pienen kuntosalin kanta-asiakkaista.

Nro	Sukupuoli	Ikä	Käyntikerrat (edellinen kuukausi)
1	nainen	29	12
2	nainen	21	16
3	mies	27	11
4	nainen	36	9
5	nainen	19	17
6	mies	25	12
7	mies	18	14
8	nainen	17	10
9	nainen	40	17
10	mies	37	23
11	nainen	33	20

12	mies	32	12
13	nainen	26	7
14	mies	20	13
15	nainen	23	11
16	mies	16	14
17	mies	20	18
18	nainen	24	15
19	mies	51	10

Avaa OpenOffice-tiedosto →

- a) Laske kaikkien kanta-asiakkaiden keski-ikä sekä erikseen miesten ja naisten keski-ivät.
- b) Kuinka monta salikäyntiä kanta-asiakkailla oli yhteensä edellisen kuukauden aikana?
- c) Havainnollista sukupuolijakaumaa ympyräkuviolla.
- d) Luokittele kanta-asiakkaat iän perusteella tasavälisiin luokkiin, joista ensimmäinen on 15–24 vuotta.
- Havainnollista luokiteltua aineistoa histogrammilla.

Ratkaisu

- a) Avataan havaintomatriisi taulukkolaskentaohjelmaan. Havaintoyksiköiden tiedot sijoittuvat omille riveilleen ja eri muuttujien arvot omiin sarakkeisiin.

	A	B	C	D
1	Nro	Sukupuoli	Ikä	Käyntikerrat (edellinen kuukausi)
2	1	nainen	29	12
3	2	nainen	21	16
4	3	mies	27	11
5	4	nainen	36	9

6	5	nainen	19	17
7	6	mies	25	12
8	7	mies	18	14
9	8	nainen	17	10
10	9	nainen	40	17
11	10	mies	37	23
12	11	nainen	33	20
13	12	mies	32	12
14	13	nainen	26	7
15	14	mies	20	13
16	15	nainen	23	11
17	16	mies	16	14
18	17	mies	20	18
19	18	nainen	24	15
20	19	mies	51	10

Lasketaan erikseen kaikkien kanta-asiakkaiden keski-ikä ja sen jälkeen miesten ja naisten keski-ikä erikseen.

Keski-ikä saadaan laskettua funktiolla KESKIARVO. Valitaan ensin tyhjä solu, johon ohjelma laskee keskiarvon. Valitaan toiminto *Lisää / Funktio*. Avautuvasta ikkunasta etsitään funktio KESKIARVO ja tuplaklikataan sitä. Kenttään *luku 1* valitaan ne solut, joiden keskiarvo lasketaan.

Klikataan OK. Kanta-asiakkaiden iän keskiarvoksi saadaan 27,1 vuotta.

Miesten ja naisten keski-iän laskemiseksi aineisto lajitellaan sukupuolen mukaan. Valitaan ensin kaikki taulukon solut ja sen jälkeen valikosta toiminto *Tiedot / Lajittele*. Lajittelu tehdään sarakkeen Sukupuoli mukaan.

Keskiarvo voidaan laskea taulukkolaskentaohjelman sijasta myös jollain muulla laskentaohjelmalla. Monet ohjelmat määrittävät samalla myös muita tunnuslukuja.

Klikataan OK. Sukupuolen mukaan lajiteltu taulukko

näyttää seuraavalaiselta:

Lajitteluavain 1

	A	B	C	D
1	Nro	Sukupuoli	Ikä	Käyntikerrat (edellinen kuukausi)
2	3	mies	27	11
3	6	mies	25	12
4	7	mies	18	14
5	10	mies	37	23
6	12	mies	32	12
7	14	mies	20	13
8	16	mies	16	14
9	17	mies	20	18
10	19	mies	51	10
11	1	nainen	29	12
12	2	nainen	21	16
13	4	nainen	36	9
14	5	nainen	19	17
15	8	nainen	17	10
16	9	nainen	40	17
17	11	nainen	33	20
18	13	nainen	26	7
19	15	nainen	23	11
20	18	nainen	24	15

☒ Nouseva
☐ Laskeva

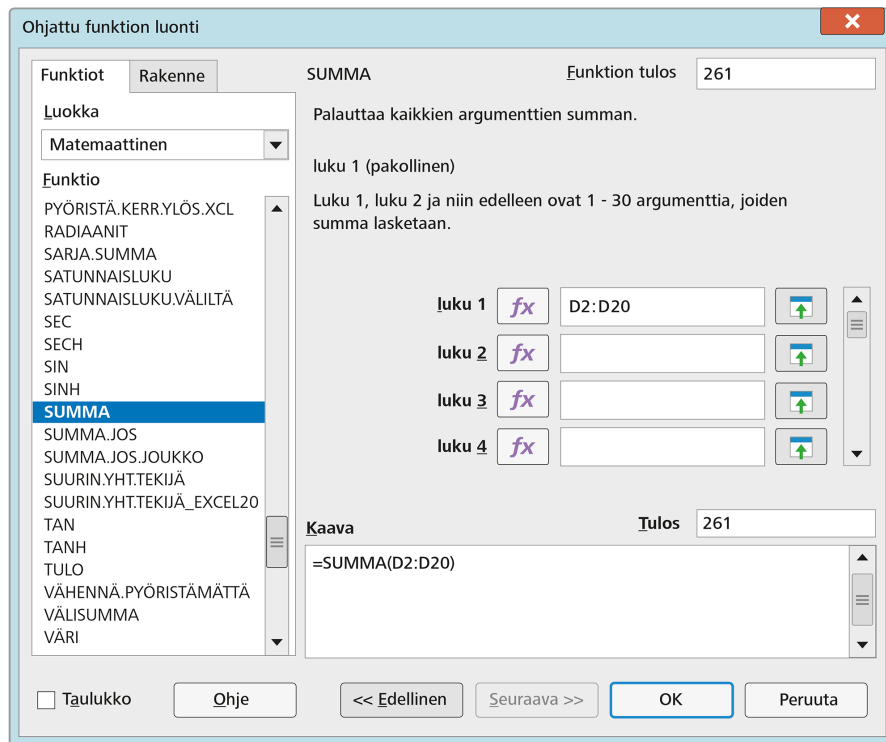
☒ Nouseva
☐ Laskeva

☒ Nouseva
☐ Laskeva

Miesten keski-ikäksi saadaan funktion KESKIARVO avulla
27,3 vuotta ja naisten keski-ikäksi 26,8 vuotta.

b) Salikäyntien yhteismäärä lasketaan taulukkolaskennan
funktioilla SUMMA. Valitaan ensin tyhjä solu, johon summa

lasketaan. Valitaan toiminto *Lisää / Funktio*. Avautuvasta ikkunasta etsitään funktio SUMMA ja tuplaklikataan sitä. Kenttään *luku 1* valitaan ne solut, joiden summa lasketaan.



Klikataan OK. Salikäynttien yhteismääräksi saadaan 261.

c) Piirretään ympyräkuvio taulukkolaskentaohjelman avulla. Ympyräkuvion piirtämistä varten miesten ja naisten lukumäärät voi poimia luokitellusta aineistosta käsin erillisiin soluihin tai luoda tietojen ohjaustaulukon, joka hoitaa ryhmittelyn ja lukumäärien laskemisen automaattisesti.

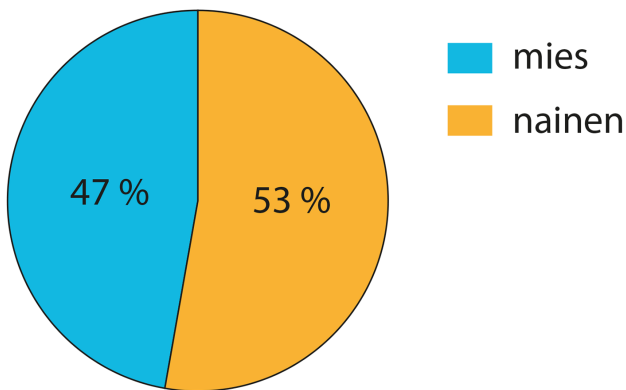
Valitaan ensin kaikki taulukon solut ja sen jälkeen valikosta toiminto *Lisää / Tietojen ohjaus*. Raahataan Sukupuoli *Rivikentät-* ja *Tietokentät-* alueisiin. Tuplaklikataan *Tietokentät-* alueen kohdetta "Summa – Sukupuoli", jolloin summan tilalle saadaan vaihdettua lukumäärä.

Klikataan OK ja uusi frekvenssitaulukko ilmestyy omalle välilehdelleen.

	A	B
1	Sukupuoli	Laske - Sukupuoli
2	mies	9
3	nainen	10
4	Yhteensä Tulos	19

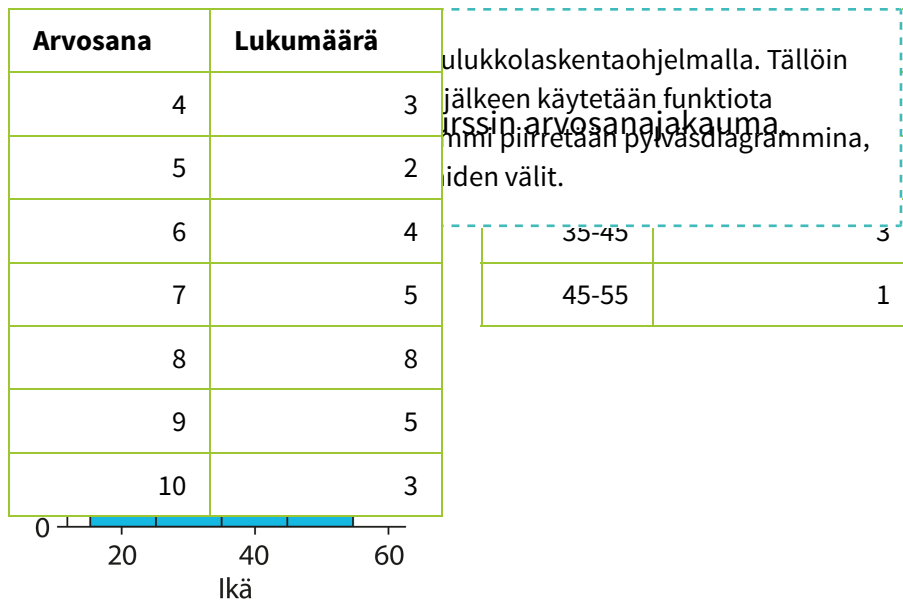
Ympyräkuvio piirretään taulukkolaskentaohjelman toiminnolla *Lisää / Kaavio*. Valitaan kaaviotyypiksi ympyrä ja lisätään kuviolle otsikko *Kaavioelementit*-kohdasta. Klikataan Valmis. Osuuksien lukumäärät tai prosenttiosuudet saadaan näkyviin valitsemalla toiminto *Lisää / Arvopisteiden otsikot*.

Kuntosalin kanta-asiakkaiden sukupuolijakauma



d) Iän mukainen luokittelu ja histogrammin piirtäminen on helpompi tehdä geometriaohjelman tilastotoiminnoilla kuin taulukkolaskentaohjelmalla.

Kopioidaan ikäsarakkeen tiedot geometriaohjelmaan ja käytetään toimintoa *Yhden muuttujan tilastot*. Ohjelma piirtää histogrammin automaattisesti. Asetuksista voidaan muokata luokkien rajoja sekä tuoda frekvenssitaulukko näkyviin. Akseleiden nimet voi lisätä, kun kopioi histogrammin ohjelman piirtoalueelle.



Avaa OpenOffice-tiedosto

→

- Laske arvosanojen keskiarvo ja keskihajonta.
- Havainnollista jakaumaa pylväskuviolla.

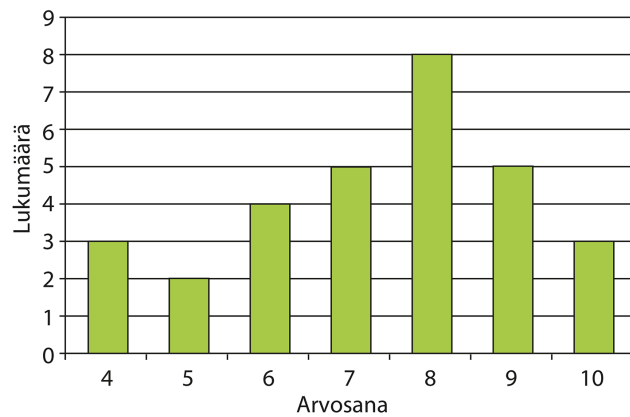
Ratkaisu

a) Taulukkolaskentaohjelmat eivät yleensä laske tilastollisia tunnuslukuja suoraan frekvenssitaulukosta. Useimmat muut laskentaohjelmat sen sijaan laskevat. Tällöin laskentaohjelman asetuksista yleensä valitaan, mistä taulukon sarakkeesta luetaan muuttujan arvot ja mistä niiden frekvenssit.

Keskiarvoksi saadaan laskentaohjelman avulla $\bar{x} \approx 7,33$ ja keskihajonnaksi $\sigma \approx 1,74$.

b) Pylväskuvio voidaan piirtää taulukkolaskentaohjelmalla. Valitaan toiminto *Lisää / Kaavio* ja kaaviotyypiksi pylväs. Jotta kuvio saadaan oikeanlaiseksi, *Tietoa*-kohdasta valitaan asetus ”Ensimmäinen sarake sisältää otsikoita”. Kuviolle voi lisätä nimen *Kaavioelementit*-kohdasta.

Ruotsin kurssin opetusryhmän
arvosanajakauma



Esimerkki 3

Taulukossa on Afrikan väkilukutietoja 1950-luvulta alkaen.

Vuosi	Väkiluku (milj. asukasta)
1955	254,0
1960	285,1
1965	322,5
1970	366,5
1975	417,9
1980	480,0
1985	552,8
1990	634,6
1995	722,9
2000	817,6
2005	924,8
2010	1 049,4
2015	1 194,4

Avaa OpenOffice-tiedosto



a) Mallinna väkiluvun kehitystä regressiosuoran avulla ja

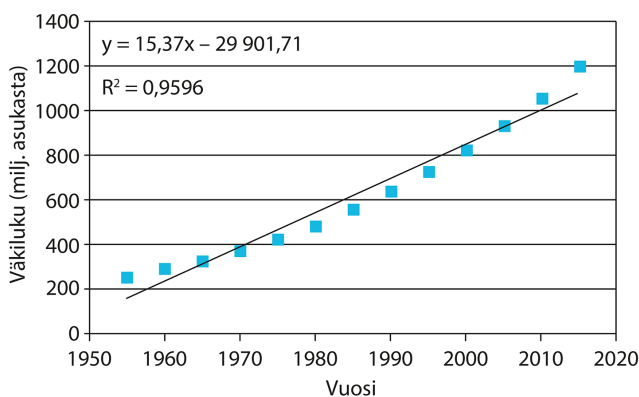
arvioi mallin sopivuutta.

b) Tutki, onko eksponentiaalinen malli tai toisen asteen polynominen malli parempi kuvaamaan väkiluvun kehitystä kuin lineaarinen malli.

Ratkaisu

a) Regressiosuoran yhtälö voidaan muodostaa esimerkiksi geometriaohjelman tai taulukkolaskentaohjelman avulla. Geometriaohjelmassa toiminto on *Kahden muuttujan regressioanalyysi*. Taulukkolaskentaohjelmassa piirretään ensin hajontakuviotoiminnolla *Lisää / Kaavio / XY (hajonta)* ja sovitetaan aineistoon trendiviiva.

Afrikan väkiluvun kehitys (lineaarinen malli)



Regressiosuoran yhtälöksi saadaan $y = 15,37x - 29\,901,71$, missä x on vuosiluku ja y on väkiluku miljoonina asukkaina.

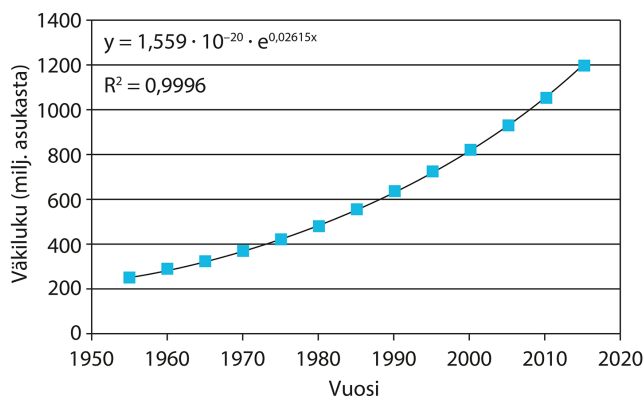
Korrelaatiokerroin on $r \approx 0,9796$, mikä viittaisi lineaarisen mallin hyvään sopivuuteen. Kuviosta kuitenkin nähdään, että pisteet sijoittuvat pikemminkin ylöspäin kaareutuvalle käyrälle kuin suoralle. Malli antaa siksi huonoja tuloksia havaintoalueen ulkopuolella eli ennen vuotta 1955 ja vuoden 2015 jälkeen.

b) Ohjelmien valikosta voidaan valita käytettävä malli. Sijoitetaan havaintoihin ensin eksponentiaalinen malli ja sitten toisen asteen polynominen malli. Mallien sopivuutta voidaan verrata selityssasteen R^2 avulla: mitä lähempänä tämä arvo on lukua 1, sitä paremmin pisteet sijoittuvat mallin mukaiselle kuvaajalle.

Lineaarisen mallin selityssaste $R^2 \approx 0,9596$ on korrelaatiokertoimen neliö.

Eksponentiaalisen mallin mukainen kuvaaja mukailee selvästi paremmin havaintopisteitä kuin lineaarinen malli.

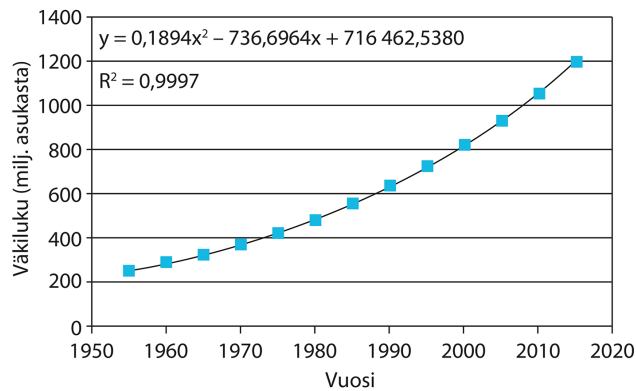
Afrikan väkiluvun kehitys (eksponentiaalinen malli)



Selityssaste mallissa on $R^2 \approx 0,9996$, joten se on parempi kuin lineaarisessa mallissa.

Toisen asteen polynomisen mallin mukainen kuvaaja mukailee havaintopisteitä myös hyvin.

Afrikan väkiluvun kehitys (toisen asteen polynominen malli)

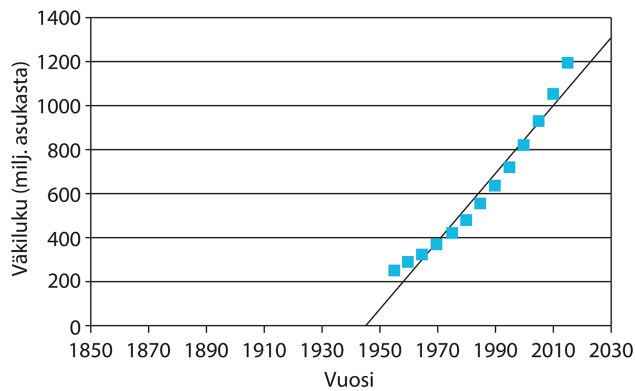


Selitysaste mallissa on $R^2 \approx 0,9997$, joten se on vielä hieman parempi kuin eksponentiaalisessa mallissa.

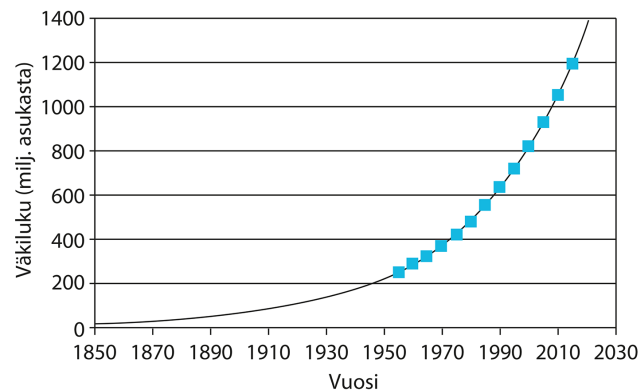
Mallin selitysaste kuvaa sen sopivuutta vain havaintopisteiden osalta. Jos havaintoalueen ulkopuolisista arvoista tiedetään jotain, malli ei saisi olla ristiriidassa tämän tiedon kanssa.

Koska toisen asteen polynomifunktion kuvaaja on paraabeli, mallin antamat tulokset alkavat muuttua nopeasti mahdottomiksi, kun mennään taaksepäin vuodesta 1955. Ohjelmasta riippuen ajanjaksoa ennen olevia arvoja voidaan tarkastella joko kopioimalla geometriaohjelmassa kuvio piirtoalueelle tai käyttämällä taulukkolaskentaohjelman toimintoa *Ekstrapoloi taaksepäin*. Tarvittaessa muutetaan myös vaaka-akselin asteikkoa. Seuraavana on esitetty kuvaajat vuosina 1850–2030.

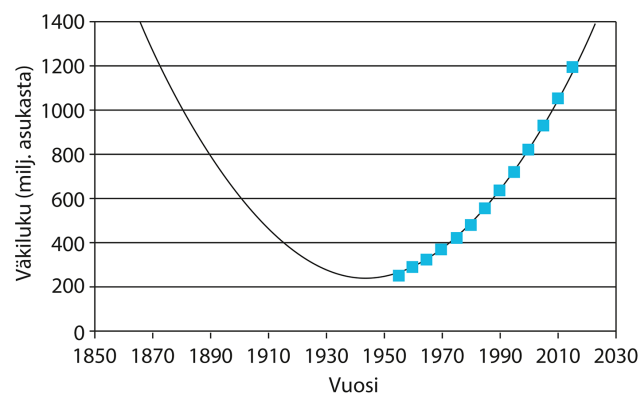
Afrikan väkiluvun kehitys (lineaarinen malli)



Afrikan väkiluvun kehitys (eksponentiaalinen malli)



Afrikan väkiluvun kehitys (toisen asteen polynominen malli)



Kuvaajista havaitaan, että lineaarinen malli antaa Afrikan väkiluvulle negatiivisia arvoja ennen vuotta 1940 ja liian pieniä arvoja vuoden 2015 jälkeen. Toisen asteen polynomisen mallin mukaan taas Afrikan väkiluku on ollut nykyistä suurempi 1800-luvulla ja sitä ennen, mikä ei pidä

paikkansa. Eksponentiaalisella mallilla on laajin pätevyysalue, joten sen voidaan sanoa olevan malleista paras.

Tietokoneohjelmat tilastoaineiston käsittelyssä

Aineiston järjestäminen jonkin ominaisuuden, kuten iän, mukaan tehdään taulukkolaskentaohjelmalla. Aineiston luokittelu esimerkiksi ikäryhmiin on helpoin tehdä geometriaohjelmassa.

Kuvioiden piirtäminen kannattaa yleensä tehdä taulukkolaskentaohjelmalla. Poikkeuksena on histogrammi, jonka saa piirrettyä nopeammin geometriaohjelmalla silloin, kun aineistoa ei ole valmiiksi luokiteltu.

Tilastolliset tunnusluvut voidaan laskea monella eri ohjelmalla. Keskiarvon ja keskihajonnan voi laskea myös peruslaskimella. Taulukkolaskentaohjelmassa funktio KESKHAJONTA antaa otoskeskihajonnan. Keskihajonnan saa funktiolla KESKHAJONTA.P.

Taulukkolaskentaohjelmalla ei saa laskettua tunnuslukuja suoraan frekvenssitaulukosta, joten tällaisessa tilanteessa kannattaa käyttää esimerkiksi geometriaohjelmaa.

Korrelaatiokertoimen ja regressiosuoran yhtälön voi määrittää esimerkiksi geometriaohjelmalla tai taulukkolaskentaohjelmalla. Geometriaohjelma määrittää molemmat yhdellä toiminnolla, mutta taulukkolaskentaohjelmassa korrelaatiokerroin pitää laskea erikseen. Muissa laskentaohjelmissa regressiosuoran yhtälö ei yleensä tule näkyviin sellaisenaan vaan kertoimet pitää poimia ohjelman antamista tunnusluvuista.

Sovitettavaksi malliksi voidaan valita myös muu kuin lineaarinen malli. Mallien sopivuutta voidaan vertailla selityssasteen R^2 avulla.