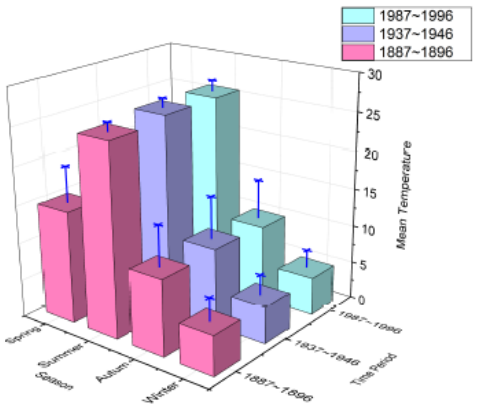
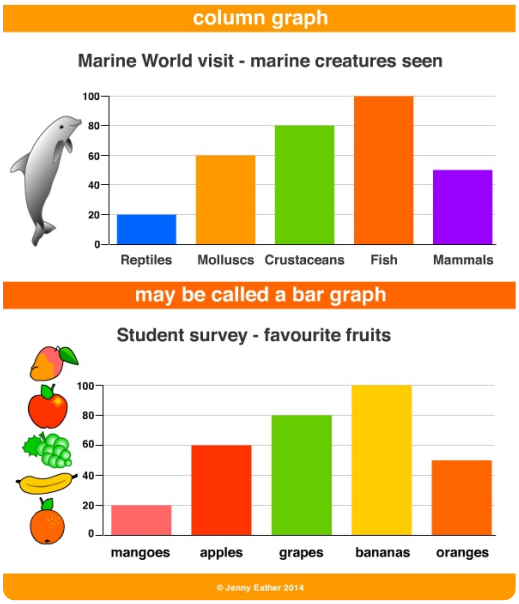
Tästä osiosta ei lasketa tehtäviä opintokorttiin, mutta osiosta tulee kolmen pisteen arvoinen välitesti.

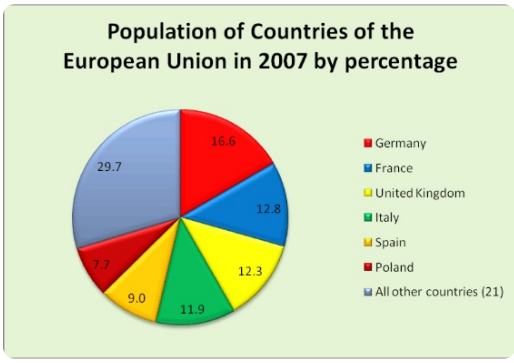
vale, emävale, tilastot

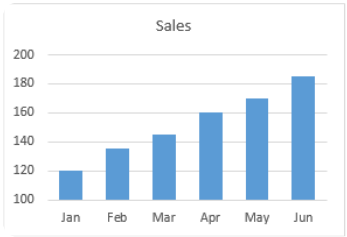
Miksi tilastot ovat niin huonossa maineessa?

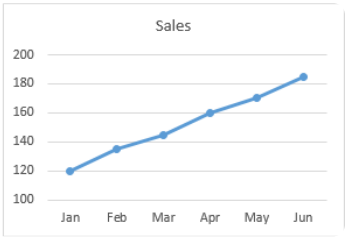
Huonoja tilastoja:











Mikä on huonoa?

- epäselvät otsakkeet ja merkinnät
- huono esitystapa (3d, piirakkadiagrammi...)
- katkaistut pylväsdiagrammit

1 Tilastojen peruskäsitteitä

Tilastotiede on numeerisen tietoaineiston keräämistä, analysointia ja tulkintaa.

- kokonaistutkimus = tutkitaan koko perusjoukkoa
- perusjoukko eli populaatio = tutkimuksen kohteena oleva koko joukko
- otantatutkimus = mitataan vain tiettyä joukkoa koko populaatiosta, mutta tehdään johtopäätelmät koko populaatioon
- edustava otos = hyvin valittu joukko, joka edustaa hyvin koko populaatiota
- harhainen otos = huonosti valittu joukko
- alkio eli tilastoyksikkö = yksittäinen tutkittava henkilö/esine/asia otoksessa
- tilastomuuttuja = alkion ominaisuus jota mitataan, esim. ihmisen pituus

- diskreetti muuttuja = saa vain yksittäisiä arvoja

Esim. Tehdään otantatutkimus, jossa mitataan puoluekannatusta. Populaationa on kaikki Suomen äänestysikäiset ja tilastomuuttujana on ihmisten mielipide äänestettävästä puolueesta. Otanta on 2000 ihmistä, jotka on valittu satunnaisesti syntymäpäivän mukaan, joten otos on edustava. Kysymys on muotoiltu näin: Jos äänestäisit tänään, mitä puoluetta äänestäisit. Koska vastausvaihtoehtojen määrä on rajallinen ja valinnaksi kelpaa vain yksi puolue (tai ei äänestä), on muuttuja diskreetti.

2. Tilastoaineksen käsittely

- frekvenssi f = ilmoittaa montako kertaa tietty (diskreetin) muuttujan arvo on toteutunut
- suhteellinen frekvenssi p = f/kaikilla (%)
- summafrekvenssi F = montako arvoa on pienempi kuin valittu x
- suhteellinen summafrekvenssi P = montako prosenttia on pienempi kuin x

Esim. Opiskelijat saivat kokeesta numerot
4,5,7,7,7,8,8,9,9,10

arvosanalle 9 f=2 ja p=2/10 = 20%

arvosanalle 9 F=7 ja P=70%

Mitta-asteikkoja on 4 erilaista

- Luokittelu (esim. puolue)
- Järjestys (esim. yo-arvosana)
- Välimatka (esim. Celsius-lämpötila)
- Suhde (esim. Kelvin lämpötila)

Huom. vaikka muuttujan arvoja kuvattaisiin numeroilla, niin silti ei voi laskea keskiarvoa, jos ei muuttuja ole vähintään välimatka-asteikkoa

Esim. Luokan syntymäkuukaudet
heinä tammi tammi helmi touko syys loka tammi joulu
1,1,1,2,5,7,9,10,12 5,33333....

Kun aineistoa on paljon, on se järkevää luokitella, kts. s. 21 esim.1

Kun tilastoituja aineistoja tutkitaan ja analysoidaan, tarvitaan *tilastollisia tunnuslukuja*:

Keskilukuja ovat mm.

- moodi eli tyyppiarvo = arvo, jonka frekvenssi on kaikkein suurin
- mediaani = havaintoarvoista keskimmäinen
- (aritmeettinen) keskiarvo

Esim. Opiskelijat saivat kokeesta numerot 4,5,7,7,7,8,8,9,9,10

Mitä mitta-asteikkoa nämä kuvastavat?

moodi = 7
mediaani =7 ja 8 tai 7,5

keskiarvo = $\frac{4+5+7+7+7+8+8+9+9+10}{10} = 7{,}4$

Keskiluvut kuvaavat, missä suurin osa arvoista on (eri tavoilla).

Varianssi ja keskihajonta ovat *hajontalukuja*, jotka kuvaavat, kuinka paljon arvot heittävät "keskeltä". Ne voidaan laskea, jos keskiarvo voidaan laskea.

Varianssi

$$\sigma^2 = \frac{(x_1-\mu)^2 + \ldots + (x_N-\mu)^2}{N} = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$$

jossa N on havaintojen lukumäärä, $x_1, \ldots, x_N$ ovat muuttujan saamat arvot ja μ on keskiarvo.

Keskihajonta σ on vain varianssin neliöjuuri.
(Standard deviation variance)

Esim. Opiskelijat saivat kokeesta numerot 4,5,7,7,7,8,8,9,9,10

Keskiarvo on 7,4. Lasketaan keskihajonta:

$$\sigma^2 = \frac{(4-7{,}4)^2 + (5-7{,}4)^2 + \ldots + (9-7{,}4)^2 + (10-7{,}4)^2}{10} = 3{,}04$$

laskin 3,377..

joten keskihajonta on $\sigma = \sqrt{3,04} \approx 1,744$

laskin 1,8...

Kun lasketaan keskihajontaa otoksista (ns. otoskeskihajonta), muuttuu kaava siten, että nimittäjässä on luvun N sijaan luku n-1.

(Hankala selitys, mutta käytännössä suurilla N arvoilla sillä ei ole merkitystä.)

Yleensä laskin laskee otoskeskihajonnan.

Välitestissä tehdään aineistosta taulukko, pylväsdiagrammi ja lasketaan sille tilastollisia tunnuslukuja.