Johdatus tilastotieteeseen

Tilastollisten aineistojen kuvaaminen

Tilastollisten aineistojen kuvaaminen

Havaintoarvojen jakauma

Tunnusluvut

Suhdeasteikollisten muuttujien tunnusluvut

Järjestysasteikollisten muuttujien tunnusluvut

Laatueroasteikollisten muuttujien tunnusluvut

Tilastollisten aineistojen kuvaaminen: Mitä opimme? – 1/2

- *Parhaan kuvan* jonkin tilastollisen muuttujan *havaittujen arvojen vaihtelusta* antaa **havaintoarvojen jakauma**.
- Jos tarkasteltava tilastollinen muuttuja on *diskreetti*, sen havainto-arvojen jakaumaa voidaan kuvata **frekvenssijakaumalla** ja sitä vastaavalla *graafisella esityksellä*, **pylväsdiagrammilla**.
- Jos tarkasteltava tilastollinen muuttuja on *jatkuva*, sen havaintoarvojen jakaumaa voidaan kuvata **luokitellulla frekvenssijakaumalla** ja sitä vastaavalla *graafisella esityksellä*, **histogrammilla**.

Tilastollisten aineistojen kuvaaminen: Mitä opimme? – 2/2

- Kuvaus havaintoarvojen jakaumasta halutaan tavallisesti *tiivistää* muutamaksi *jakauman karakteristisia ominaisuuksia kuvaavaksi* **tunnusluvuksi**.
- *Keskimääräisten*, *tyypillisten* tai *yleisten havaintoarvojen sijaintia* kuvataan **keskiluvuilla**.
- Havaintoarvojen *keskittymistä* tai *hajaantumista* jonkin keskiluvun ympärillä kuvataan **hajontaluvuilla**.
- Myös havaintoarvojen jakauman *vinoutta* ja *huipukkuutta* voidaan kuvata sopivasti valituilla tunnusluvuilla.
- Tarkasteltavan tilastollisen muuttujan *mitta-asteikolliset ominaisuudet määräävät, mitä tunnuslukuja muuttujaa koskevista havaintoarvoista saa ja kannattaa laskea*.

Tilastollisten aineistojen kuvaaminen: Esitiedot

- Esitiedot: ks. seuraavaa lukua:

  Tilastollisten aineistojen kerääminen ja mittaaminen

Tilastollisten aineistojen kuvaaminen: Lisätiedot

- Tilastollisia aineistoja kuvaavien tunnuslukujen *otosjakaumia* käsitellään luvussa

  Otos ja otosjakaumat

Tilastollisten aineistojen kuvaaminen

>> Havaintoarvojen jakauma

Tunnusluvut

Suhdeasteikollisten muuttujien tunnusluvut

Järjestysasteikollisten muuttujien tunnusluvut

Laatueroasteikollisten muuttujien tunnusluvut

Havaintoarvojen jakauma

Avainsanat

- **Frekvenssijakauma**
- **Frekvenssit**
- **Havaintoarvojen jakauma**
- **Havaintoarvot**
- **Histogrammi**
- **Luokiteltu frekvenssijakauma**
- **Luokkafrekvenssit**
- **Mitta-asteikot ja havaintoarvojen jakauman kuvaaminen**
- **Pylväsdiagrammi**
- **Tilastolliset aineistot**

Tilastollinen aineisto

- Tilastollisen tutkimuksen *kaikki mahdolliset kohteet* muodostavat tutkimuksen (**kohde-**) **perusjoukon**.
- Tutkimuksen kohteiksi valittuja perusjoukon alkioita kutsutaan **havaintoyksiköiksi**.
- **Tilastollinen aineisto** koostuu havaintoyksiköitä kuvaavien muuttujien **havaituista arvoista**.
- Huomautuksia:
  - Tilastollinen aineisto voi syntyä *tilastollisen kokeen* tuloksena tai tekemällä *suoria havaintoja*.
  - Jos tutkimuksen kohteena on koko perusjoukko, tutkimusta kutsutaan *kokonaistutkimukseksi*, muuten kyseessä on *otantatutkimus*.

Havaintoarvot

- Olkoon tutkimuksen kohteiksi valittujen **havainto-yksiköiden lukumäärä** n.
- Olkoon

  $$x_i \, , \; i = 1, 2, \ldots , n$$

  kohdeperusjoukon alkioiden ominaisuutta kuvaavan muuttujan x **havaittu arvo** havaintoyksikössä i.
- Kutsumme muuttujan x havaittuja arvoja

  $$x_1 \, , x_2 \, , \ldots , x_n$$

  tavallisesti **havaintoarvoiksi** tai **havainnoiksi**.
- Havaintoarvo x_i saadaan *mittaamalla* muuttujan x arvo havaintoyksikölle i.

Havaintoarvojen jakauma ja sen kuvaaminen 1/4

- Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan *x* *havaittujen arvojen*

  $$x_1 , x_2 , \ldots , x_n$$

  vaihtelua havaintoyksiköiden joukossa kuvaa parhaiten havaintoarvojen **jakauma**.

Havaintoarvojen jakauma ja sen kuvaaminen 2/4

- Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan *x* havaittujen arvojen

  $$x_1 , x_2 , \ldots , x_n$$

  jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:

  – Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valitulla **graafisella esityksellä**.

  – *Jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **tunnusluvuilla**.

Havaintoarvojen jakauma ja sen kuvaaminen 3/4

- Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan x (mitta-asteikolliset) ominaisuudet (ks. lukua **Tilastollisten aineistojen kerääminen ja mittaaminen**) määräävät muuttujan x havaittujen arvojen

  $$x_1 , x_2 , \ldots , x_n$$

  jakaumalle parhaiten sopivan kuvaustavan; ks. seuraavaa kalvoa.

Havaintoarvojen jakauma ja sen kuvaaminen 4/4

- Jos muuttuja *x* on *diskreetti*, sen havaittujen arvojen jakaumaa voidaan kuvata **frekvenssijakaumalla** ja sitä vastaavalla graafisella esityksellä **pylväsdiagrammilla**.
- Jos muuttuja *x* on *jatkuva*, sen havaittujen arvojen jakaumaa voidaan kuvata **luokitellulla frekvenssi-jakaumalla** ja sitä vastaavalla graafisella esityksellä **histogrammilla**.

Frekvenssit

- Olkoon muuttuja *x diskreetti* ja olkoot

  $y_1 , y_2 , \ldots , y_m$

  muuttujan *x mahdolliset arvot*.
- Olkoot

  $x_1 , x_2 , \ldots , x_n$

  muuttujan *x havaitut arvot*.
- Muuttujan *x mahdollisen arvon* y_k , $k = 1, 2, \ldots , m$ **frekvenssi**

  f_k

  kertoo *kuinka monta kertaa* y_k esiintyy havaintoarvojen $x_1 , x_2 , \ldots , x_n$ joukossa.

Frekvenssijakauma

- Muuttujan *x mahdolliset arvot*

  $y_1 , y_2 , \cdots , y_m$

  yhdessä niiden *frekvenssien*

  $f_1 , f_2 , \cdots , f_m$

  kanssa muodostavat muuttujan *x havaittujen arvojen*

  $x_1 , x_2 , \cdots , x_n$

  frekvenssijakauman.
- Huomaa, että

  $f_1 + f_2 + \cdots + f_m = n$

  jossa *n* on havaintojen kokonaislukumäärä.

Pylväsdiagrammi

- *Frekvenssijakaumaa*

  $(y_k , f_k) , k = 1, 2, \ldots , m$

  voidaan kuvata graafisesti **pylväsdiagrammilla**, jossa muuttujan *x mahdollisen arvon* y_k *frekvenssiä* f_k havaintoarvojen $x_1 , x_2 , \ldots , x_n$ joukossa esittää pisteeseen y_k piirretty *pylväs, jonka korkeus vastaa frekvenssiä* f_k .
- Huomautus:

  Pylväsdiagrammin tulkinta on analoginen *diskreetin todennäköisyysjakauman pistetodennäköisyysfunktion* tulkinnan kanssa; ks. lukua **Satunnaismuuttujat ja todennäköisyysjakaumat.**

Pylväsdiagrammin piirtäminen: Havainnollistus 1/2

- Olkoot

  $y_1 , y_2 , \ldots , y_m$

  muuttujan x mahdolliset arvot ja olkoon

  (y_k , f_k)

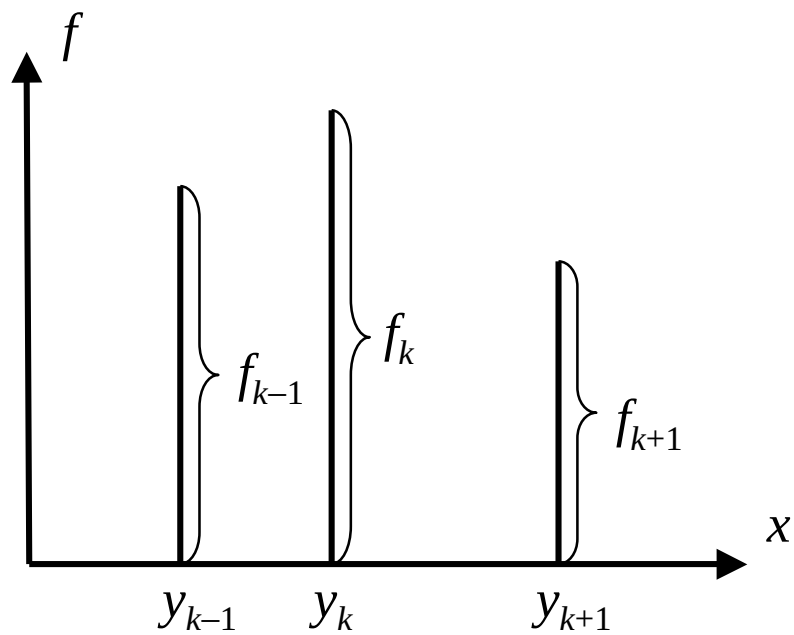
  $k = 1, 2, \ldots , m$

  muuttujan x havaittujen arvojen

  $x_1 , x_2 , \ldots , x_n$

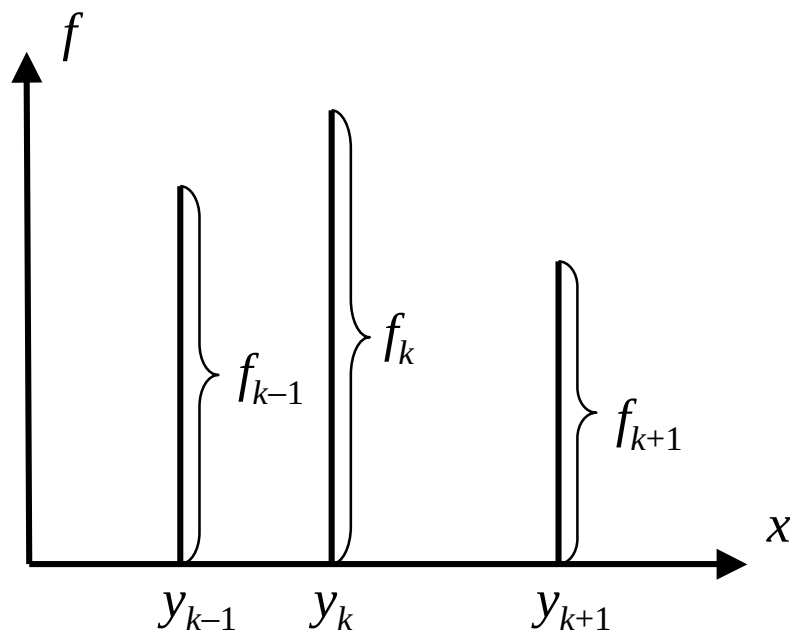
  frekvenssijakauma.

- *Frekvenssi* f_k kertoo kuinka monta kertaa muuttujan x arvo y_k esiintyy havaintoarvojen joukossa.



Pylväsdiagrammin piirtäminen: Havainnollistus 2/2

- Tarkastellaan muuttujan x mahdollista arvoa y_k vastaavan *pylvään* piirtämistä pylväs-diagrammiin.
- Muuttujan x mahdolliset arvot y_k määräävät pylväiden *paikat*.
- Pylvään *korkeus* valitaan suhteessa arvon y_k frekvenssiin f_k .



Pylväsdiagrammi: Esimerkki 1/2

- Matemaattisen tilastotieteen kurssille osallistui 20 opiskelijaa.
- Kurssin loppukokeen tehtävän 4 arvosteluasteikkona oli 0-6 pistettä niin, että

  0 = huonoin pistemäärä

  6 = paras pistemäärä
- Opiskelijoiden saamat pisteet on annettu ylemmässä taulukossa oikealla.
- Alemmassa taulukossa on annettu pisteiden *frekvenssi-jakauma*.

Pisteet; *n* = 20

0	0	0	0	0
0	1	1	1	2
5	5	5	5	5
6	6	6	6	6

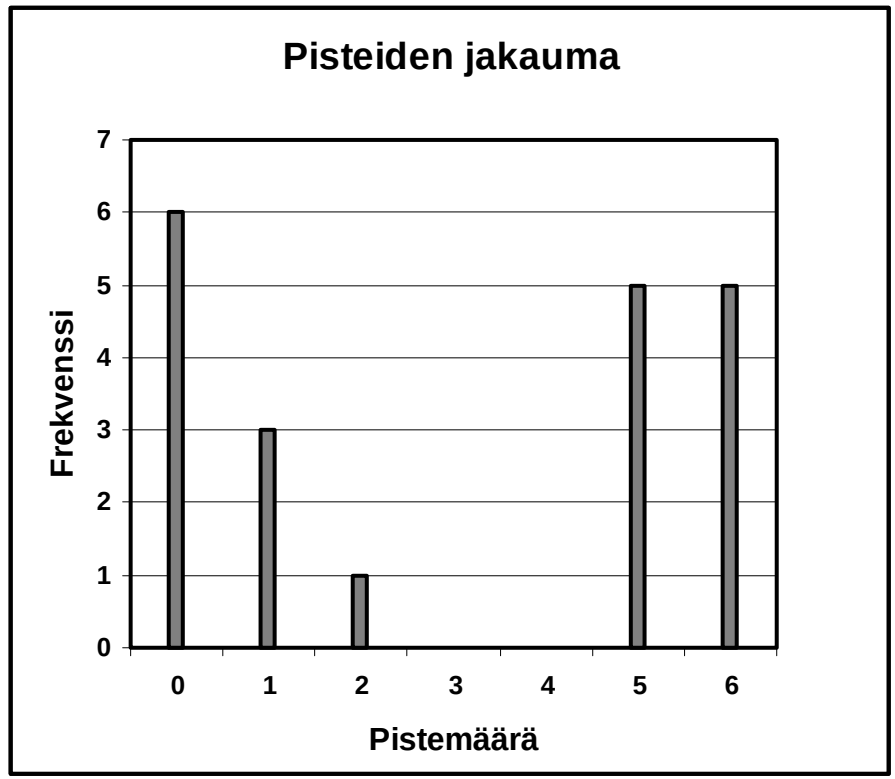
Pisteet	Frekvenssi
0	6
1	3
2	1
3	0
4	0
5	5
6	5

Pylväsdiagrammi: Esimerkki 2/2

- Kuva oikealla esittää pisteiden frekvenssijakaumaa vastaavaa *pylväsdiagrammia*.
- Muuttujan

  x = pistemäärä

  mahdolliset arvot määräävät pylväiden *paikan*.
- Pylväät on piirretty niin, että niiden *korkeudet* vastaavat muuttujan x mahdollisten arvojen *frekvenssejä*.



Luokkafrekvenssit 1/2

- Olkoon muuttuja *x jatkuva* ja oletetaan, että sen *mahdolliset arvot* ovat *välillä*

  (a, b)

  jossa voi olla $a = -\infty$, $b = +\infty$.
- Jaetaan väli (a, b) pisteillä

  $$a = a_0 < a_1 < a_2 < \quad \cdots \quad < a_{m-1} < a_m = b$$

  pistevieraisiin *osaväleihin*

  $$(a_{k-1}\ ,\ a_k]\ ,\ k = 1, 2, \ldots, m$$

Luokkafrekvenssit 2/2

- Olkoot

  $x_1 , x_2 , \ldots , x_n$

  muuttujan *x havaitut arvot*.
- Muuttujan *x* havaittujen arvojen **frekvenssi**

  f_k

  luokassa *k* kertoo niiden havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *lukumäärän*, jotka kuuluvat väliin

  $(a_{k-1} , a_k] , k = 1, 2, \ldots , m$

Luokiteltu frekvenssijakauma

- *Luokkavälit*

  $$(a_{k-1}\,,\, a_k]\,,\; k = 1, 2, \ldots, m$$

  yhdessä vastaavien *luokkafrekvenssien*

  $$f_1\,,\, f_2\,,\, \ldots\,,\, f_m$$

  kanssa muodostavat muuttujan *x havaittujen arvojen*

  $$x_1\,,\, x_2\,,\, \ldots\,,\, x_n$$

  luokitellun frekvenssijakauman.

- Huomaa, että

  $$f_1 + f_2 + \cdots + f_m = n$$

  jossa *n* on havaintojen kokonaislukumäärä.

Histogrammi

- *Luokiteltua frekvenssijakaumaa*

  $$((a_{k-1}\ ,\ a_k]\ ,\ f_k)\ ,\ k = 1,\ 2,\ \ldots\ ,\ m$$

  voidaan kuvata graafisesti **histogrammilla**, jossa muuttujan x havaittujen arvojen $x_1\ ,\ x_2\ ,\ \ldots\ ,\ x_n$ frekvenssiä f_k luokassa $(a_{k-1}\ ,\ a_k]$ esittää *suorakaide*, jonka *kantana* on väli

  $$(a_{k-1}\ ,\ a_k]$$

  ja jonka *pinta-ala vastaa luokkafrekvenssiä* f_k .

- Huomautus:

  Histogrammin tulkinta on analoginen *jatkuvan todennäköisyys-jakauman tiheysfunktion* tulkinnan kanssa; ks. lukua **Satunnaismuuttujat ja todennäköisyysjakaumat**.

Histogrammin piirtäminen: Havainnollistus 1/2

- Olkoon

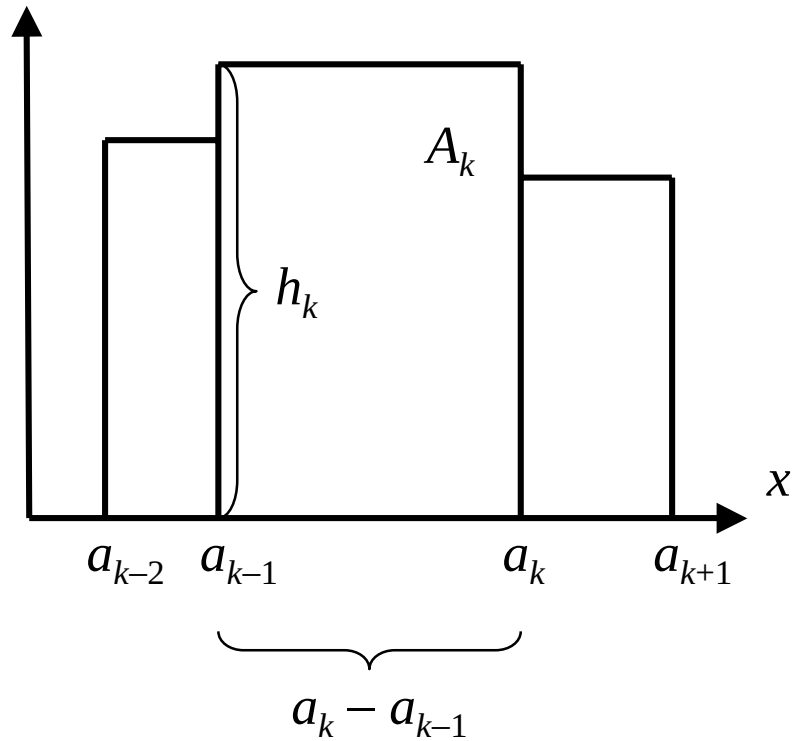
  $((a_{k-1}\,,\, a_k]\,,\, f_k)$

  $k = 1, 2, \ldots, m$

  muuttujan x havaittujen arvojen

  $x_1\,,\, x_2\,,\, \ldots\,,\, x_n$

  luokiteltu frekvenssijakauma.
- *Luokkafrekvenssi* f_k kertoo niiden havaintoarvojen lukumäärän, jotka kuuluvat *luokkaväliin* $(a_{k-1}\,,\, a_k]$.



Histogrammin piirtäminen: Havainnollistus 2/2

- Tarkastellaan *k*. luokkaa vastaavan *suorakaiteen* piirtämistä histogrammiin.
- Luokkaväli

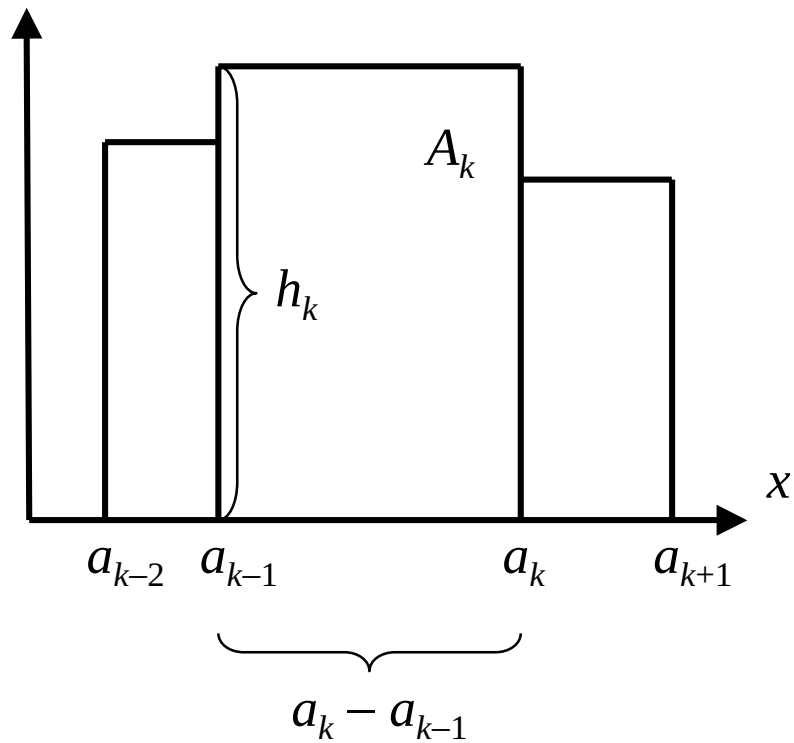
  $(a_{k-1}\ ,\ a_k]$

  muodostaa suorakaiteen *kannan*.
- Suorakaiteen *korkeus* h_k saadaan ehdosta

  A_k = *k*. luokkaa vastaavan suorakaiteen pinta-ala

  $= (a_k - a_{k-1}) \times h_k$

  $= f_k$



Histogrammi: Esimerkki 1/3

- Kone tekee *ruuveja*, joiden *pituudet vaihtelevat satunnaisesti*.
- Poimitaan ruuvien joukosta *yksinkertainen satunnaisotos*, jonka *koko*

  $n = 30$

  ja mitataan otokseen poimittujen ruuvien pituudet.
- Otokseen poimittujen 30:n ruuvin pituudet (yksikkö = cm) on annettu oikealla olevassa taulukossa.

Ruuvien pituudet; *n* = 30

10.05	10.23	10.02	10.24	10.14
10.06	10.07	10.09	10.00	10.09
10.30	10.17	10.18	10.00	10.01
10.00	9.93	10.16	10.21	10.20
9.99	10.13	9.88	9.99	10.12
10.20	9.93	10.00	10.07	10.13

Histogrammi: Esimerkki 2/3

- Muodostetaan otokseen poimittujen ruuvien pituuksien *luokiteltu frekvenssijakauma*.
- Järjestetään sitä varten havainto-arvot *suuruusjärjestykseen*; ks. ylempää taulukkoa oikealla.
- Pituuksien *luokiteltu frekvenssi-jakauma* on annettu alemmassa taulukossa.
- Esimerkiksi luokkaan, jonka määrää puoliavoin väli

  (10.10, 10.15]

  kuuluu 4 ruuvia.

Ruuvien pituudet; *n* = 30

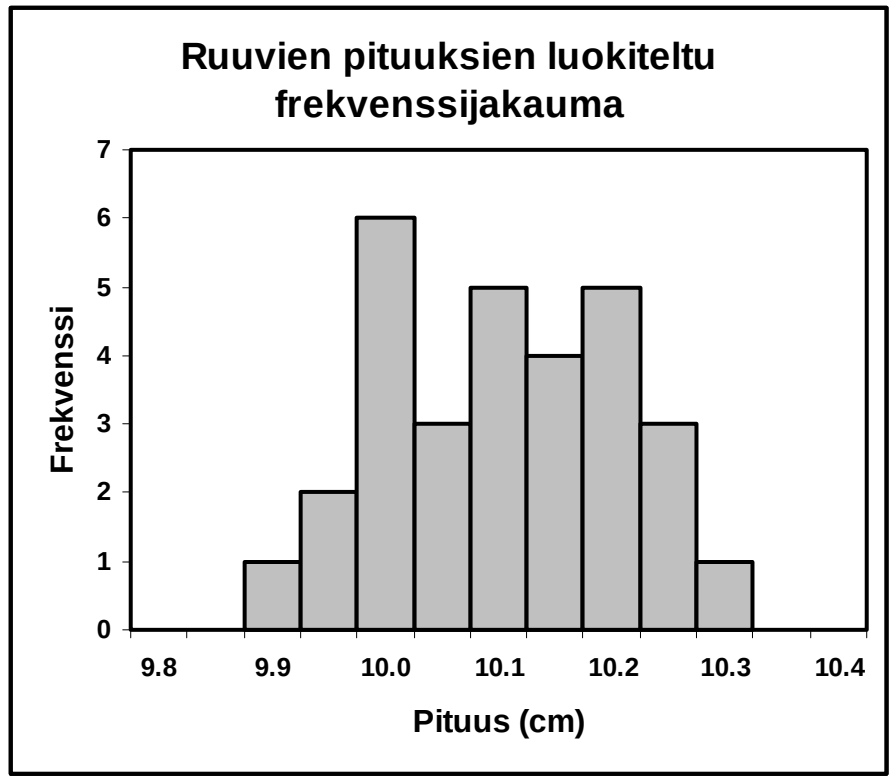
9.88	9.93	9.93	9.99	9.99
10.00	10.00	10.00	10.00	10.01
10.02	10.05	10.06	10.07	10.07
10.09	10.09	10.12	10.13	10.13
10.14	10.16	10.17	10.18	10.20
10.20	10.21	10.23	10.24	10.30

Luokkavälit	Luokkafrekvenssit
(9.85,9.90]	1
(9.90,9.95]	2
(9.95,10.00]	6
(10.00,10.05]	3
(10.05,10.10]	5
(10.10,10.15]	4
(10.15,10.20]	5
(10.20,10.25]	3
(10.25,10.30]	1

(c) Ilkka Mellin (2005)

Histogrammi: Esimerkki 3/3

- Kuva oikealla esittää otokseen poimittujen ruuvien pituuksien luokiteltua frekvenssijakaumaa vastaavaa *histogrammia*.
- *Luokkavälit* määräävät histogrammin suorakaiteiden *kannat*.
- Suorakaiteet on piirretty niin, että niiden *pinta-alat* vastaavat *luokkafrekvenssejä*.



Mitta-asteikot ja havaintoarvojen jakauman kuvaaminen

- **Laatuero- tai järjestysasteikollisten muuttujien** havaittujen arvojen kuvaamiseen käytettävät välineet:
  - **Frekvenssijakauma**
  - **Pylväsdiagrammi**
- **Välimatka- tai suhdeasteikollisten muuttujien** havaittujen arvojen kuvaamiseen käytettävät välineet:
  - **Luokiteltu frekvenssijakauma**
  - **Histogrammi**

Mitta-asteikot: ks. lukua **Tilastollisten aineistojen kerääminen ja mittaaminen**.

Tilastollisten aineistojen kuvaaminen

- Havaintoarvojen jakauma
- >> Tunnusluvut
- Suhdeasteikollisten muuttujien tunnusluvut
- Järjestysasteikollisten muuttujien tunnusluvut
- Laatueroasteikollisten muuttujien tunnusluvut

Tunnusluvut

Avainsanat

- **Mitta-asteikot ja niille sallitut tunnusluvut**
- **Tunnusluvut havaintoaineiston kuvaajina**
- **Tunnusluvut ja mitta-asteikot**

Tunnusluvut havaintoaineiston kuvaajina 1/4

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  muuttujan *x havaittuja arvoja.*
- Muuttujan *x* havaittujen arvojen jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:
  - *Jakaumaa kokonaisuutena* voidaan kuvata sopivasti valitulla **graafisella esityksellä**.
  - *Jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **tunnusluvuilla**.

Tunnusluvut havaintoaineiston kuvaajina 2/4

- Tunnuslukujen tehtävänä on kuvata havaintoarvojen jakauman keskeisiä *karakteristisia ominaisuuksia*:
  - *Keskimääräisten*, *tyypillisten* tai *yleisten* havainto-arvojen *sijaintia* kuvataan **keskiluvuilla**.
  - Havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* kuvataan **hajontaluvuilla**.
  - Myös havaintoarvojen jakauman **vinoutta** ja **huipukkuutta** voidaan kuvata sopivasti valituilla tunnusluvuilla.

Tunnusluvut havaintoaineiston kuvaajina 3/4

- Havaintoarvojen jakauman *karakteristisia ominaisuuksia* on syytä tavallisesti kuvata usealla erilaisella tunnusluvulla.
- Havaintoaineiston *jakauma* ja *kuvauksen tavoitteet* määräävät mitä tunnuslukuja havaintoaineistosta *kannattaa* laskea.
- Tutkittavan muuttujan *mitta-asteikolliset ominaisuudet* määräävät mitä tunnuslukuja havaintoaineistosta *saa* laskea.

Tunnusluvut havaintoaineiston kuvaajina 4/4

- Huomautuksia:
  - Tunnuslukujen antama kuvaus havaintoarvojen jakaumasta jää *puutteelliseksi* ja saattaa olla jopa *harhaanjohtava*, ellei sitä *täydennetä* sopivilla jakaumaa kuvaavilla *graafisilla esityksillä* kuten *pylväsdiagrammilla* tai *histogrammilla*.
  - Havaintoarvojen jakaumaa on tavallisesti syytä kuvata *usealla eri tavalla*.

Tunnusluvut ja mitta-asteikot

- Tarkasteltavan muuttujan *mitta-asteikolliset ominaisuudet ohjaavat* havaintoaineiston kuvaamisessa käytettävien *tunnuslukujen valintaa*.

  Mitta-asteikot: ks. lukua **Tilastollisten aineistojen kerääminen ja mittaaminen**.

- Tunnusluvut voidaan ryhmitellä tarkastelun kohteena olevien muuttujien mitta-asteikollisten ominaisuuksien perusteella seuraavalla tavalla:
  - **Tunnusluvut välimatka- ja suhdeasteikollisille muuttujille**
  - **Tunnusluvut järjestysasteikollisille muuttujille**
  - **Tunnusluvut laatueroasteikollisille muuttujille**

Välimatka- ja suhdeasteikollisten muuttujien tunnuslukuja

- Tunnuslukuja **välimatka- ja suhdeasteikollisten muuttujien** havaituille arvoille:
  - **Aritmeettinen keskiarvo** keskilukuna
  - **Varianssi** ja **keskihajonta** hajontalukuina
  - **Origomomentit**
  - **Keskusmomentit**
  - **Vinous**
  - **Huipukkuus**
  - **Harmoninen keskiarvo**
  - **Geometrinen keskiarvo**

Järjestysasteikollisten muuttujien tunnuslukuja

- Tunnuslukuja **järjestysasteikollisten muuttujien** havaituille arvoille:
  - **Järjestystunnusluvut**
  - **Mimimi** ja **maksimi**
  - **Vaihteluväli** ja **vaihteluvälin pituus**
  - **Prosenttipisteet**
  - **Mediaani** keskilukuna
  - **Kvartiilit**
  - **Kvartiiliväli** ja **kvartiilivälin pituus**
  - **Kvartiilipoikkeama** hajontalukuna

Laatueroasteikollisten muuttujien tunnuslukuja

- Tunnuslukuja **laatueroasteikollisten muuttujien** havaituille arvoille:
  - **Suhteellinen frekvenssi**
  - **Moodi** keskilukuna

Mitta-asteikot ja niille sallitut tunnusluvut 1/3

- **Välimatka- ja suhdeasteikollisille muuttujille** sallitut tunnusluvut:
  - **Origo-** ja **keskusmomentit** ja niistä johdetut tunnusluvut
  - Kaikki *laatuero-* ja *järjestysasteikollisten muuttujien tunnusluvut*
  - *Keskilukuna* käytetään tavallisesti **aritmeettista keskiarvoa**, mutta monissa tilanteissa keskilukuna on syytä käyttää **mediaania** tai **moodia**
  - *Hajontalukuna* käytetään tavallisesti **keskihajontaa** tai **varianssia**

Mitta-asteikot ja niille sallitut tunnusluvut 2/3

- **Järjestysasteikollisille muuttujille** sallitut tunnusluvut:
  - **Järjestystunnusluvut** ja niistä johdetut tunnusluvut
  - Kaikki *laatueroasteikollisten muuttujien tunnusluvut*
  - *Keskilukuna* käytetään tavallisesti **mediaania**, mutta monissa tilanteissa keskilukuna on syytä käyttää **moodia**
  - *Hajontalukuna* käytetään usein **kvartiilipoikkeamaa**
- Huomautus:

  Välimatka- tai *suhdeasteikollisten* muuttujien tunnuslukuja *ei ole mielekästä laskea järjestysasteikollisten muuttujien havaituille arvoille.*

Mitta-asteikot ja niille sallitut tunnusluvut 3/3

- **Laatueroasteikollisille muuttujille** sallitut tunnusluvut:
  - **Suhteelliset frekvenssit**
  - *Keskilukuna* käytetään **moodia**
- Huomautus:

  Järjestys-, *välimatka-* tai *suhdeasteikollisten* muuttujien tunnuslukuja *ei ole mielekästä laskea laatueroasteikollisten muuttujien havaituille arvoille.*

Tilastollisten aineistojen kuvaaminen

- Havaintoarvojen jakauma
- Tunnusluvut
- >> Suhdeasteikollisten muuttujien tunnusluvut
- Järjestysasteikollisten muuttujien tunnusluvut
- Laatueroasteikollisten muuttujien tunnusluvut

Suhdeasteikollisten muuttujien tunnusluvut

Avainsanat

- **Aritmeettinen keskiarvo**
- **Geometrinen keskiarvo**
- **Harmoninen keskiarvo**
- **Huipukkuus**
- **Keskihajonta**
- **Keskusmomentit**
- **Luokitellun aineiston aritmeettinen keskiarvo**
- **Origomomentit**
- **Standardointi**
- **Tilastollinen etäisyys**
- **Varianssi**
- **Vinous**

Tunnusluvut suhdeasteikollisille muuttujille

- Tavallisimmat tunnusluvut *suhdeasteikollisten* muuttujien havaituille arvoille:
  - **Aritmeettinen keskiarvo** keskilukuna
  - **Varianssi** ja **keskihajonta** hajontalukuina
  - **Origomomentit**
  - **Keskusmomentit**
  - **Vinous**
  - **Huipukkuus**
  - **Harmoninen keskiarvo**
  - **Geometrinen keskiarvo**

Aritmeettinen keskiarvo

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja.
- **Aritmeettinen keskiarvo**

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

  kuvaa havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *keskimääräistä* arvoa.
- Aritmeettisesta keskiarvosta (engl. *mean*) käytetään usein myös symbolia *M*.

Luokitellun aineiston aritmeettinen keskiarvo

- Oletetaan, että *jatkuvan* muuttujan *x* havaituista arvoista on muodostettu *luokiteltu frekvenssijakauma* ja olkoon käytetty *luokkien lukumäärä k*.
- Oletetaan, että *luokkakeskuksina* ovat luvut

  $$z_1 , z_2 , \ldots , z_k$$

  ja että vastaavat *luokkafrekvenssit* ovat

  $$f_1 , f_2 , \ldots , f_k$$
- Tällöin **luokitellun aineiston aritmeettinen keskiarvo** on

  $$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} f_i z_i$$

  jossa $n = \sum f_i$.

Aritmeettinen keskiarvo jakauman kuvaajana

- Aritmeettinen keskiarvo kuvaa havaintoarvojen *keski-määräistä* arvoa.
- Havaintoarvojen aritmeettinen keskiarvo sijoittuu havaintoarvojen jakauman *painopisteeseen*.
- Jos havaintoarvojen jakauma on *vino* tai *monihuippuinen*, aritmeettinen keskiarvo *ei välttämättä ole tyypillinen* tai *yleinen havaintoarvo*.
- Aritmeettinen keskiarvo *ei ole robusti* eli se *on herkkä poikkeaville havaintoarvoille*, koska *jokainen havainto-arvo vetää aritmeettista keskiarvoa puoleensa*; ks. havainnollistusta seuraavalla kalvolla.

Aritmeettisen keskiarvon herkkyys poikkeaville havainnoille

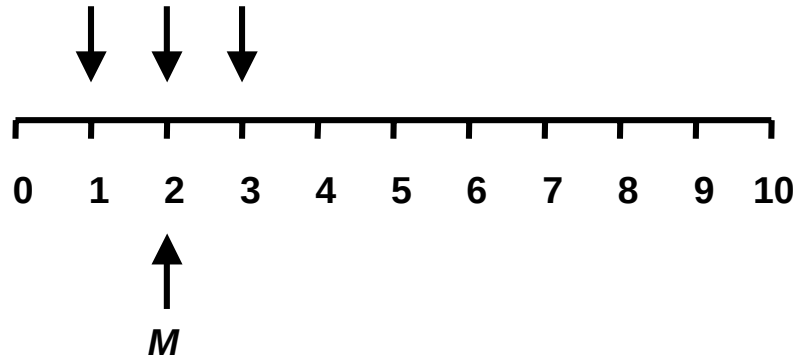
- Aritmeettinen keskiarvo on *herkkä* poikkeaville havainnoille.
- Havaintoarvojen 1, 2, 3 aritmeettinen keskiarvo on

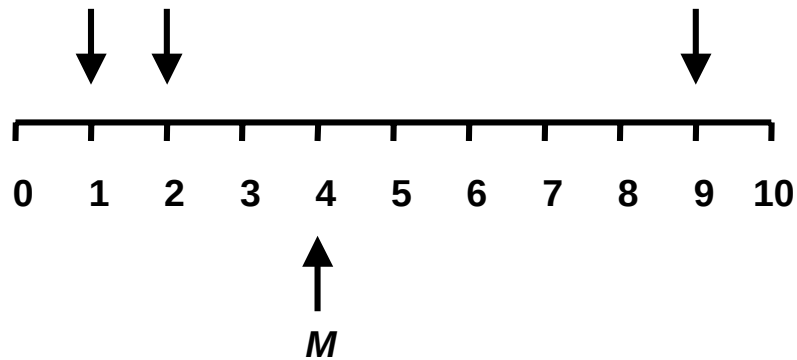
$$M = \frac{1+2+3}{3} = 2$$

- Muutetaan havaintoarvo 3 havaintoarvoksi 9 ja *pidetään muut havaintoarvot samoina*.
- Tällöin *uudeksi* aritmeettiseksi keskiarvoksi tulee

$$M = \frac{1+2+9}{3} = 4$$

- Ks. kuvaa oikealla.





Varianssi 1/2

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja ja olkoon $\bar{x}$ havaintoarvojen *aritmeettinen keskiarvo*.

- (**Otos-**) **varianssi**

  $$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

  kuvaa havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *hajaantuneisuutta* tai *keskittyneisyyttä* niiden painopisteen $\bar{x}$ ympärillä.

Varianssi 2/2

- Havaintoarvojen x_1 , x_2 , … , x_n otosvarianssi lasketaan usein myös kaavalla

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

  jossa summalausekkeen jakajana on *n*.

- Huomautus:

  Otosvarianssin kaksi erilaista kaavaa liittyvät erilaisiin tapoihin *estimoida* normaalijakauman N(μ, σ^2) *varianssiparametri* σ^2 :

  (i) s^2 on *harhaton estimaattori* parametrille σ^2 .

  (ii) $\hat{\sigma}^2$ on parametrin σ^2 *suurimman uskottavuuden estimaattori*.

Varianssi: Toinen laskukaava

- Jos otosvarianssi joudutaan laskemaan *käsin* tai *laskimella* havaintoarvojen x_1 , x_2 , … , x_n *varianssi kannattaa laskea* kaavalla

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right)$$

tai vaihtoehtoisen kaavan tapauksessa kaavalla

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right)$$

Varianssi: Toisen laskukaavan todistus 1/2

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja ja olkoon

  $$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  havaintoarvojen *aritmeettinen keskiarvo*.

Varianssi: Toisen laskukaavan todistus 2/2

- Tällöin

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}^2) \\
&= \sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\overline{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-2\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i+n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 \\
&= \left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)
\end{aligned}
$$

Keskihajonta 1/2

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja ja olkoon $\bar{x}$ havaintoarvojen *aritmeettinen keskiarvo*.

- (**Otos-**) **keskihajonta**

  $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

  on otosvarianssin s^2 neliöjuuri ja kuvaa havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *hajaantuneisuutta* tai *keskittyneisyyttä* niiden painopisteen $\bar{x}$ ympärillä.

Keskihajonta 2/2

- Havaintoarvojen x_1 , x_2 , … , x_n (**otos-**) **keskihajonta** lasketaan usein myös kaavalla

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

  jossa summalausekkeen jakajana on *n*.

- Huomautus:

  Keskihajonnan kaksi erilaista kaavaa liittyvät erilaisiin tapoihin *estimoida* normaalijakauman N(μ, σ^2) *varianssiparametri* σ^2 :

  (i) s^2 on *harhaton estimaattori* parametrille σ^2 .

  (ii) $\hat{\sigma}^2$ on parametrin σ^2 *suurimman uskottavuuden estimaattori*.

Keskihajonta ja varianssi jakauman kuvaajina 1/2

- Keskihajonta ja varianssi ovat *havaintoarvojen vaihtelun mittoja*.
- *Varianssi* on havaintoarvojen keskimääräinen neliöllinen poikkeama niiden aritmeettisesta keskiarvosta.
- Havaintoarvojen *keskihajonta* on varianssin neliöjuuri.
- Jos havaintoarvojen jakaumaa kuvaavana *keskilukuna* on käytetty *aritmeettista keskiarvoa*, *hajontalukuna* on luontevaa käyttää *keskihajontaa*:
  - (i) Keskihajonnalla ja aritmeettisella keskiarvolla *on sama dimensio* (*laatu*).
  - (ii) Varianssin ja aritmeettisen keskiarvon *dimensio* (*laatu*) *ei ole sama*.

Keskihajonta ja varianssi jakauman kuvaajina 2/2

- ”*Pieni*” keskihajonta (varianssi) merkitsee sitä, että havaintoarvot *keskittyvät* niiden painopisteen (aritmeettisen keskiarvon) ympärille.
- ”*Suuri*” keskihajonta (varianssi) merkitsee sitä, että havaintoarvot *ovat hajaantuneet* niiden painopisteen (aritmeettisen keskiarvon) ympärille.
- Varianssi ja keskihajonta eivät ole *robusteja* eli ne ovat *herkkiä poikkeaville havaintoarvoille*.

Aritmeettinen keskiarvo ja varianssi: Laskutoimitusten suorittaminen 1/2

- Oletetaan, että haluamme laskea havaintoarvojen

  $$x_1 , x_2 , \ldots , x_n$$

  aritmeettisen keskiarvon $\bar{x}$ ja *otosvarianssin* s^2 *käsin* tai käyttämällä *laskinta*
- Tällöin tarvittavat laskutoimitukset on mukavinta järjestää seuraavalla kalvolla esitettävän kaavion muotoon.

Aritmeettinen keskiarvo ja varianssi: Laskutoimitusten suorittaminen 2/2

- Havaintoarvojen *aritmeettinen keskiarvo* ja *varianssi* voidaan laskea määräämällä ensin havaintoarvojen *summa* ja *neliösumma* sekä käyttämällä sen jälkeen alla esitettyjä kaavoja.

i	x_i	x_i^2
1	x_1	x_1^2
2	x_2	x_2^2
$\vdots$	$\vdots$	$\vdots$
n	x_n	x_n^2
Summa	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

Standardointi

- Olkoot $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen x_1 , x_2 , … , x_n *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*.
- Tällöin **standardoitujen havaintoarvojen**

$$z_i = \frac{x_i - \bar{x}}{s_x}, i = 1, 2, \ldots, n$$

aritmeettinen keskiarvo ja varianssi ovat

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = 0$$

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \bar{z})^2 = 1$$

Tilastollinen etäisyys

- Olkoot $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen $x_1 , x_2 , \ldots , x_n$ *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*.
- Havaintoarvojen x_k ja x_l **tilastollinen etäisyys** d_{kl} on

$$d_{kl} = \frac{x_k - x_l}{s_x}$$

- Havaintoarvojen x_k ja x_l tilastollinen etäisyys *ottaa* etäisyyttä määrättäessä *huomioon kaikkien havainto-arvojen* $x_1 , x_2 , \ldots , x_n$ *vaihtelun*.
- Huomautus:

  Tilastollisessa testauksessa käytettävät *testisuureet* voidaan usein tulkita *tilastollisen etäisyyden mittareiksi*; ks. lukuja **Testit ...** .

Origomomentit

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja.
- Havaintoarvojen $x_1 , x_2 , \ldots , x_n$ ***k.* origomomentti** on

  $$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \; , k = 1, 2, 3, \ldots$$
- Erityisesti 1. origomomentti a_1 on havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *aritmeettinen keskiarvo*:

  $$a_1 = \overline{x}$$

Keskusmomentit

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja ja olkoon $\bar{x}$ havaintoarvojen *aritmeettinen keskiarvo*.

- Havaintoarvojen $x_1 , x_2 , \ldots , x_n$ ***k*. keskusmomentti** on

  $$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \ , k = 1,2,3,\ldots$$

- Erityisesti $m_1 = 0$ kaikille havaintoarvoille ja

  $$m_2 = \hat{\sigma}^2 = a_2 - a_1^2$$

  on havaintoarvojen $x_1 , x_2 , \ldots , x_n$ *varianssi*.

Vinous

- Olkoot

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

havaintoarvojen

$$x_1 , x_2 , \ldots , x_n$$

2. ja vastaavasti 3. *keskusmomentti*.

- Tunnuslukua

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

käytetään kuvaamaan havaintoarvojen jakauman **vinoutta**.

Vinous jakauman kuvaajana 1/3

- Jos havaintoarvojen jakauma on *symmetrinen painopisteensä suhteen*,

  $c_1 \approx 0$

- Esimerkki:

  Normaalijakautuneilla havaintoaineistoilla $c_1 \approx 0$.

Vinous jakauman kuvaajana 2/3

- Jos

  $$c_1 > 0$$

  sanomme, että havaintoarvojen jakauma on **positiivisesti vino**.
- Oletetaan, että $c_1 > 0$ ja havaintoarvojen jakaumaa kuvaava *pylväsdiagrammi* (diskreetin muuttujan tapauksessa) tai *histogrammi* (jatkuvan muuttujan tapauksessa) on *yksihuippuinen.*
- Tällöin jakaumaa kuvaava diagrammi on **vino oikealle** eli sen oikeanpuoleinen häntä on pitempi kuin sen vasemmanpuoleinen häntä.

Vinous jakauman kuvaajana 3/3

- Jos

  $$c_1 < 0$$

  sanomme, että havaintoarvojen jakauma on **negatiivisesti vino**.
- Oletetaan, että $c_1 < 0$ ja havaintoarvojen jakaumaa kuvaava *pylväsdiagrammi* (diskreetin muuttujan tapauksessa) tai *histogrammi* (jatkuvan muuttujan tapauksessa) on *yksihuippuinen.*
- Tällöin jakaumaa kuvaava diagrammi on **vino vasemmalle** eli sen vasemmanpuoleinen häntä on pitempi kuin sen oikeanpuoleinen häntä.

Huipukkuus

- Olkoot

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

havaintoarvojen

$$x_1 , x_2 , \ldots , x_n$$

2. ja vastaavasti 4. *keskusmomentti*.

- Tunnuslukua

$$c_2 = \frac{m_4}{m_2^2} - 3$$

käytetään kuvaamaan havaintoarvojen jakauman **huipukkuutta**.

Huipukkuus jakauman kuvaajana

- *Normaalijakautuneilla* havaintoaineistoilla $c_2 \approx 0$.
- Olkoon havaintoarvojen jakauman huipukkuus

  $$c_2 > 0$$

  Tällöin jakauma on **huipukas** (normaalijakautuneeseen havaintoaineistoon verrattuna).
- Olkoon havaintoarvojen jakauman huipukkuus

  $$c_2 < 0$$

  Tällöin jakauma on **laakea** (normaalijakautuneeseen havaintoaineistoon verrattuna).

Harmoninen keskiarvo

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  positiivisen välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja.
- Havaintoarvojen $x_1 , x_2 , \ldots , x_n$ **harmoninen keskiarvo** on

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Harmoninen keskiarvo: Esimerkki 1/2

- Esimerkki osoittaa, että *aritmeettinen keskiarvo ei ole kaikissa tilanteissa sopiva keskiluku*.
- Olkoon kahden kaupungin *A* ja *B* välimatka 120 km.
- Ajetaan matka *A*:sta *B*:hen 60 km/h ja matka *B*:stä *A*:han 120 km/h.
- Mikä on ollut *keskinopeus* edestakaisella matkalla?

Matka *A*:sta *B*:hen ja takaisin	=	240 km
Ajoaika *A*:sta *B*:hen	=	2 h
Ajoaika *B*:stä *A*:han	=	1 h
Ajoaika yhteensä	=	3 h
Keskinopeus edestakaisella matkalla	=	240/3 = 80 km/h

Harmoninen keskiarvo: Esimerkki 2/2

- Nopeuksien *aritmeettinen keskiarvo*

$$M = \frac{60 + 120}{2} = 90 \text{ km/h}$$

antaa *väärän* keskinopeuden.

- Sen sijaan nopeuksien *harmoninen keskiarvo*

$$H = \frac{1}{\frac{1}{2}\left(\frac{1}{60} + \frac{1}{120}\right)} = 80 \text{ km/h}$$

antaa *oikean* keskinopeuden.

Geometrinen keskiarvo

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  positiivisen välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja.
- Havaintoarvojen $x_1 , x_2 , \ldots , x_n$ **geometrinen keskiarvo** on

  $$G = \sqrt[n]{x_1 x_2 \cdots x_n}$$
- Huomautus:

  Geometrisen keskiarvon logaritmi on havaintoarvojen logaritmien aritmeettinen keskiarvo:

  $$\log(G) = \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_n)}{n}$$

Geometrinen keskiarvo: Esimerkki 1/4

- Esimerkki osoittaa, että *aritmeettinen keskiarvo ei ole kaikissa tilanteissa sopiva keskiluku*.
- Olkoon lainan suuruus 100 € .
- Olkoon korkoprosentti 1. vuotena 10 % ja 2. vuotena 20 % .
- Jos lainaa ei lyhennetä, lainapääoma karttuu seuraavalla tavalla:

Pääoma 1. vuoden lopussa = $1.1 \times 100 = 110$ €

Pääoma 2. vuoden lopussa = $1.2 \times 110 = 132$ €

- Lainapääoma karttuu siis kahdessa vuodessa 32 % .
- Jos kumpanakin vuotena käytettäisiin *samaa korkoprosenttia*, miten se pitäisi valita, jotta lainapääoma olisi 2. vuoden lopussa 132 € ?

Geometrinen keskiarvo: Esimerkki 2/4

- Korkoprosenttien *aritmeettinen keskiarvo*

$$M = \frac{10 + 20}{2} = 15\ \%$$

tuottaa *väärän* lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = $1.15 \times 100 = 115$ €

Pääoma 2. vuoden lopussa = $1.15 \times 115 = 132.25$ €

Geometrinen keskiarvo: Esimerkki 3/4

- Korkoprosentti

$$\frac{32}{2} = 16\ \%$$

tuottaa *väärän* lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = $1.16 \times 100 = 116$ €

Pääoma 2. vuoden lopussa = $1.16 \times 116 = 134.56$ €

Geometrinen keskiarvo: Esimerkki 4/4

- Sen sijaan *geometrinen keskiarvo*

$$G = \sqrt{1.1 \times 1.2} = 1.1489125$$

antaa korkoprosentiksi

14.89125 %

joka tuottaa *oikean* lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = $1.1489125 \times 100 = 114.89125$ €

Pääoma 2. vuoden lopussa = $1.1489125 \times 114.89125$

= 132.00 €

Aritmeettinen, harmoninen ja geometrinen keskiarvo

- Oletetaan, että *aritmeettinen keskiarvo M*, *harmoninen keskiarvo H* ja *geometrinen keskiarvo G* määrätään *samoista* positiivisista luvuista $x_1 , x_2 , \ldots , x_n$.
- Tällöin

$$H \leq G \leq M$$

ja

$$H = G = M$$

jos ja vain jos

$$x_1 = x_2 = \cdots = x_n$$

Tilastollisten aineistojen kuvaaminen

- Havaintoarvojen jakauma
- Tunnusluvut
- Suhdeasteikollisten muuttujien tunnusluvut
- >> Järjestysasteikollisten muuttujien tunnusluvut
- Laatueroasteikollisten muuttujien tunnusluvut

Järjestysasteikollisten muuttujien tunnusluvut

Avainsanat

- **Box and Whisker -kuvio**
- **Järjestystunnusluvut**
- **Kvartiilipoikkeama**
- **Kvartiilit**
- **Kvartiiliväli ja kvartiilivälin pituus**
- **Luokitellun aineiston mediaani**
- **Mediaani**
- **Mimimi ja maksimi**
- **Prosenttipisteet**
- **Robustisuus**
- **Vaihteluväli ja vaihteluvälin pituus**

Tunnusluvut järjestysasteikollisille muuttujille 1/2

- Tavallisimmat tunnusluvut *järjestysasteikollisten* muuttujien havaituille arvoille:
  - **Järjestystunnusluvut**
  - **Mimimi** ja **maksimi**
  - **Vaihteluväli** ja **vaihteluvälin pituus**
  - **Prosenttipisteet**
  - **Mediaani** keskilukuna
  - **Kvartiilit**
  - **Kvartiiliväli** ja **kvartiilivälin pituus**
  - **Kvartiilipoikkeama** hajontalukuna

Tunnusluvut järjestysasteikollisille muuttujille 2/2

- Havaintoaineistojen jakaumia voidaan usein havainnollistaa kätevästi **Box and Whisker -kuviolla**.
- Huomautus:

  Järjestysasteikollisten muuttujien tunnuslukuja *saa käyttää ja on usein myös järkevää käyttää kuvaamaan välimatka-* ja *suhde-asteikollisten muuttujien havaittujen arvojen jakaumaa*.

Järjestystunnusluvut

- Olkoot

  $$x_1 , x_2 , \ldots , x_n$$

  järjestys-, *välimatka-* tai *suhdeasteikollisen* muuttujan x havaittuja arvoja.
- *Järjestetään* havaintoarvot $x_1 , x_2 , \ldots , x_n$ suuruus-järjestykseen pienimmästä suurimpaan ja olkoot

  $$z_1 , z_2 , \ldots , z_n$$

  järjestykseen asetetut havaintoarvot.
- Suuruusjärjestyksessä k. havaintoarvoa z_k kutsutaan ***k*. järjestystunnusluvuksi**.

Minimi, maksimi ja vaihteluväli

- Olkoot

  $$z_1 , z_2 , \ldots , z_n$$

  järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

- Tällöin

  z_1 = **minimiarvo**

  z_n = **maksimiarvo**

  (z_1 , z_n) = **vaihteluväli**

  $z_n - z_1$ = **vaihteluvälin pituus**

Prosenttipisteet

- Olkoot

  $$z_1 , z_2 , \ldots , z_n$$

  järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

- Havaintoarvojen ***p*. prosenttipiste**

  $$z_{(p)} , p = 1, 2, \ldots , 99$$

  on piste, joka jakaa havaintoaineiston *kahteen osaan*:

  (i) p % havaintoarvoista on lukua $z_{(p)}$ *pienempiä* tai korkeintaan yhtä suuria kuin $z_{(p)}$.

  (ii) $(100 - p)$ % havaintoarvoista on lukua $z_{(p)}$ *suurempia*.

Mediaani

- Olkoot

  $$z_1 , z_2 , \ldots , z_n$$

  järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.
- **Mediaani** *Me* on havaintoarvojen 50. prosenttipiste:

  $$Me = z_{(50)}$$
- Mediaani jakaa havaintoaineiston *kahteen yhtä suureen osaan* niin, että toisessa *kaikki* havaintoarvot ovat mediaania *pienempiä*, toisessa *kaikki* havaintoarvot ovat mediaania *suurempia*.

Mediaanin laskeminen

- Havaintoarvojen mediaani *Me* voidaan määrätä seuraavalla tavalla:

  (1) Järjestetään havaintoarvot *suuruusjärjestykseen* pienimmästä suurimpaan.

  (2a) Jos havaintoarvojen lukumäärä on *pariton*, mediaani on järjestetyistä havaintoarvoista *keskimmäinen*.

  (2b) Jos havaintoarvojen lukumäärä on *parillinen*, mediaani on järjestetyistä havaintoarvoista *kahden keskimmäisen aritmeettinen keskiarvo*.

Luokitellun aineiston mediaani

- **Luokitellun aineiston mediaani** voidaan laskea kaavalla

$$Me = L_i + \frac{\frac{1}{2}n - \Sigma f_j}{f_i} \times c_i$$

jossa

L_i = mediaaniluokan alaraja

$\sum f_j$ = kaikkien mediaaniluokan alapuolella oleviin luokkiin kuuluvien havaintoarvojen frekvenssi

f_i = mediaaniluokkaan kuuluvien havaintoarvojen frekvenssi

c_i = mediaaniluokan pituus

n = havaintoarvojen lukumäärä

Mediaani jakauman kuvaajana 1/2

- Mediaani on suuruusjärjestykseen asetettujen havainto-arvojen *keskimmäinen* havaintoarvo (tai kahden keskimmäisen aritmeettinen keskiarvo).
- Jos havaintoarvojen jakauma on *symmetrinen*, havaintoarvojen mediaani ja aritmeettinen keskiarvo yhtyvät.
- Jos havaintoarvojen jakauma on *vino*, mutta yksi-huippuinen, havaintoarvojen mediaani kuvaa *tyypillisiä* havaintoarvoja usein paremmin kuin niiden aritmeettinen keskiarvo.
- Jos havaintoarvojen jakauma on *monihuippuinen*, mediaani *ei välttämättä ole yleinen havaintoarvo*.

Mediaani jakauman kuvaajana 2/2

- Mediaani on **robusti** eli se *ei ole* – toisin kuin aritmeettinen keskiarvo – *herkkä poikkeaville havaintoarvoille*.

Mediaanin robustisuus: Havainnollistus

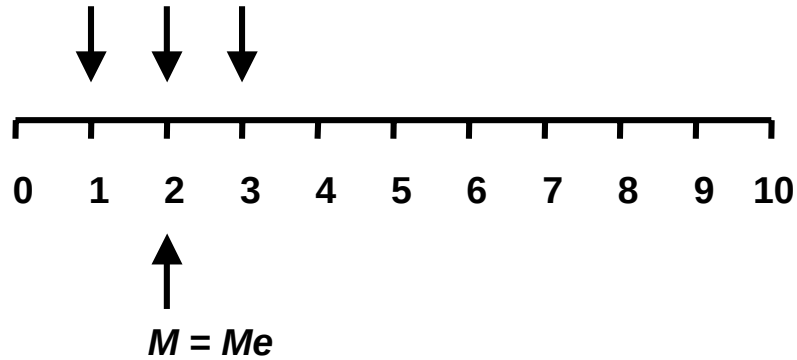
- Havaintoarvojen 1, 2, 3 aritmeettinen keskiarvo on

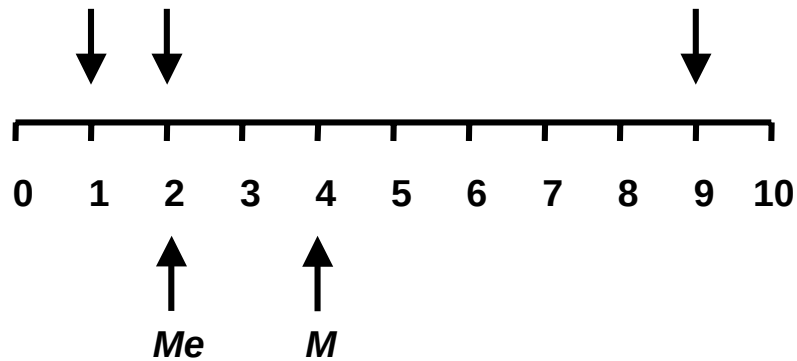
$$M = \frac{1+2+3}{3} = 2$$

- Muutetaan havaintoarvo 3 havaintoarvoksi 9 ja *pidetään muut havaintoarvot samoina*.
- Tällöin *uudeksi* aritmeettiseksi keskiarvoksi tulee

$$M = \frac{1+2+9}{3} = 4$$

- Sen sijaan havaintoarvojen mediaani *Me ei muutu*.
- Ks. kuvaa oikealla.





Mediaani, aritmeettinen keskiarvo ja jakauman vinous

- Oletetaan, että *aritmeettinen keskiarvo M* ja *mediaani Me* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen *luokitellusta frekvenssijakaumasta*.
- Jos havaintoarvojen jakauma on *yksihuippuinen*, pätee seuraava (ks. havainnollistusta seuraavalla kalvolla):

  (i) *Vasemmalle vinoilla jakaumilla*

  $$M < Me$$

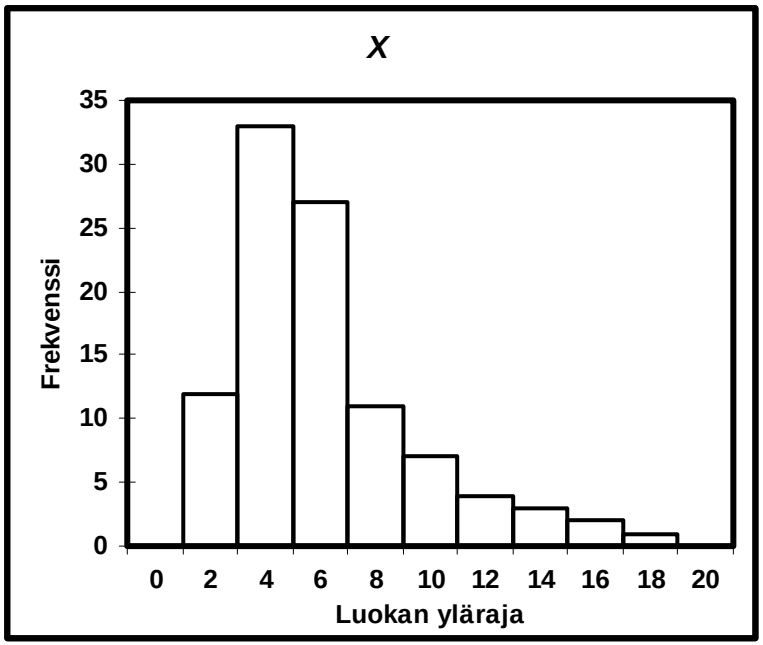
  (ii) *Symmetrisillä jakaumilla*

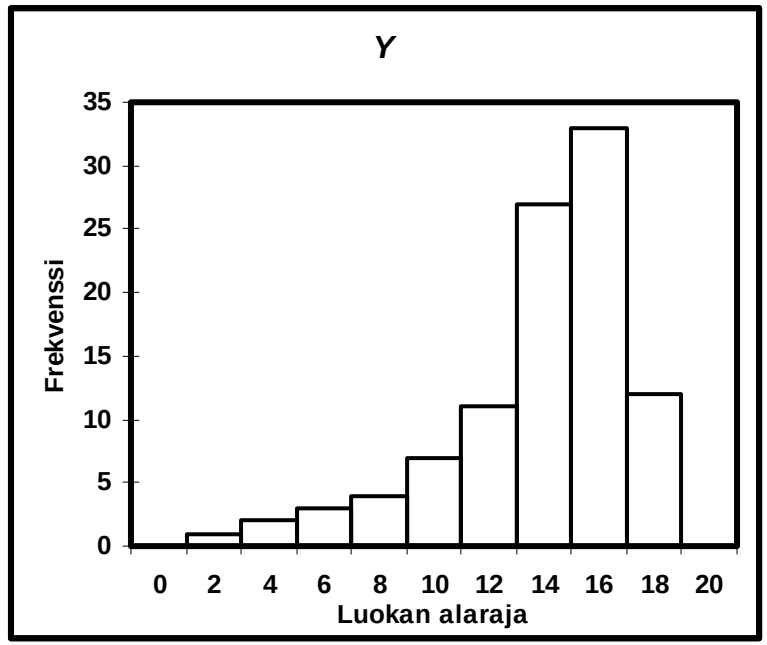
  $$M \approx Me$$

  (iii) *Oikealle vinoilla jakaumilla*

  $$Me < M$$

Mediaani, aritmeettinen keskiarvo ja jakauman vinous: Havainnollistus 1/2



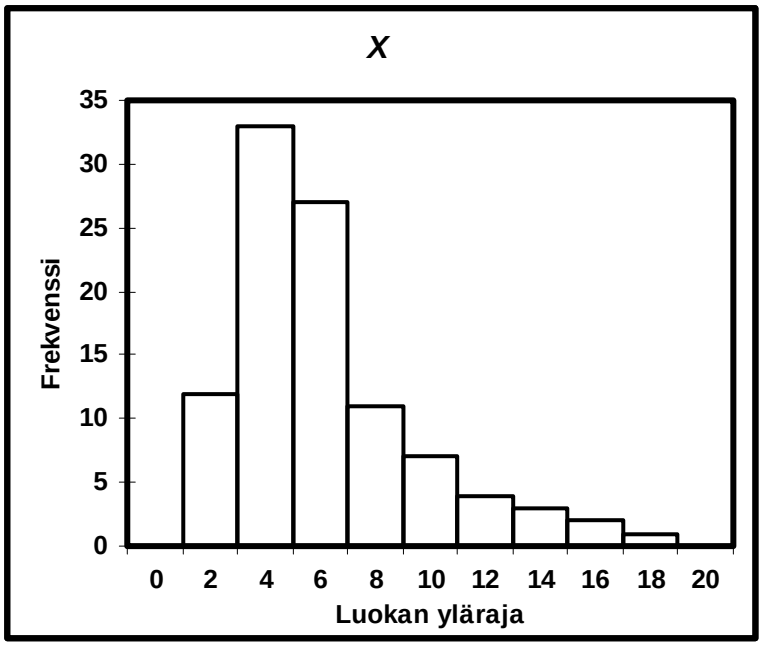


- Yllä olevat histogrammit perustuvat sataan satunnaislukujen avulla generoituun havaintoarvoon:

$$X \sim \chi^2(5)$$

$$Y = 20 - X$$

Mediaani, aritmeettinen keskiarvo ja jakauman vinous: Havainnollistus 2/2

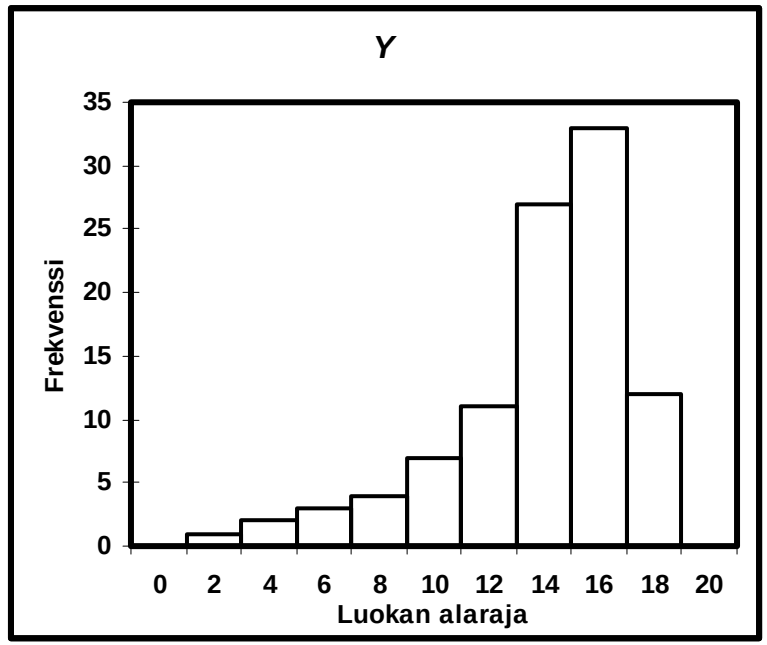


- Jakauma on vino *oikealle*:

  Vinous = 1.25

  Aritmeettinen keskiarvo = 5.19

  Mediaani = 4.41



- Jakauma on vino *vasemmalle*:

  Vinous = –1.25

  Aritmeettinen keskiarvo = 14.81

  Mediaani = 15.59

Kvartiilit 1/2

- Olkoot

  $$z_1 , z_2 , \ldots , z_n$$

  järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

- Tällöin

  Q_1 = **Alakvartiili** = 25. prosenttipiste = $z_{(25)}$

  Q_2 = **Keskikvartiili** = 50. prosenttipiste = $z_{(50)}$

  Q_3 = **Yläkvartiili** = 75. prosenttipiste = $z_{(75)}$

Kvartiilit 2/2

- Kvartiilit Q_1 , Q_2 , Q_3 jakavat suuruusjärjestykseen asetetun havaintoaineiston *neljään yhtä suureen osaan*.
- Erityisesti:

  Keskikvartiili Q_2

  = Havaintoarvojen *mediaani Me*

  Alakvartiili Q_1

  = Havaintoarvojen mediaania *Me pienempien havaintoarvojen mediaani*

  Yläkvartiili Q_3

  = Havaintoarvojen mediaania *Me suurempien havaintoarvojen mediaani*

Kvartiilit, kvartiiliväli ja kvartiilipoikkeama

- Olkoot havaintoarvojen *kvartiilit* Q_1 , Q_2 , Q_3 .
- Tällöin

(Q_1 , Q_3) = **kvartiiliväli**

$Q_3 - Q_1 = IQR$ = **kvartiilivälin pituus**

$(Q_3 - Q_1)/2 = IQR/2$ = **kvartiilipoikkeama**

- Kvartiiliväliä, kvartiilivälin pituutta (*IQR* = interquartile range) ja kvartiilipoikkeamaa voidaan käyttää kuvaamaan havaintoarvojen *hajaantuneisuutta* (*keskittyneisyyttä*).
- Jos havaintoarvojen jakaumaa kuvaavana *keskilukuna* on käytetty *mediaania*, *hajontalukuna* käytetään usein *kvartiilipoikkeamaa*.

Box and Whisker -kuvio 1/6

- Havaintoarvojen jakaumaa voidaan usein kätevästi havainnollistaa ns. **Box and Whisker -kuviolla**.
- Olkoon

  (Q_1 , Q_3)

  havaintoarvojen *kvartiiliväli* ja

  $Me = Q_2$

  havaintoarvojen *mediaani.*
- Olkoon

  $IQR = Q_3 - Q_1$

  havaintoarvojen *kvartiilivälin* (Q_1 , Q_3) *pituus*.

Box and Whisker -kuvio 2/6

- Määritellään kuvion *sisäaidat* f_1 ja f_3 kaavoilla

  $$f_1 = Q_1 - 1.5 \times IQR$$

  $$f_3 = Q_3 + 1.5 \times IQR$$

- Olkoon a_1 *pienin* havaintoarvo, joka toteuttaa ehdon

  $$a_1 \geq f_1$$

- Olkoon a_3 *suurin* havaintoarvo, joka toteuttaa ehdon

  $$a_3 \leq f_3$$

- Määritellään kuvion *ulkoaidat* F_1 ja F_3 kaavoilla

  $$F_1 = Q_1 - 3 \times IQR$$

  $$F_3 = Q_3 + 3 \times IQR$$

Box and Whisker -kuvio 3/6

- *Box and Whisker -kuvio* koostuu *laatikosta*, *viiksistä* ja kuvioon merkityistä *poikkeuksellisista havainnoista*.
- Box and Whisker -kuvion piirtäminen:

  (i) Piirretään suorakaiteen muotoinen **laatikko** kuvaamaan havaintoarvojen *kvartiiliväliä*

  (Q_1 , Q_3)

  Merkitään havaintoarvojen *mediaani*

  $Me = Q_2$

  laatikkoon poikkiviivalla.

Box and Whisker -kuvio 4/6

Box and Whisker -kuvion piirtäminen (jatkuu):

(ii) Piirretään jananmuotoiset **viikset** laatikon molemmille puolille kuvaamaan välejä

(a_1 , Q_1) ja (Q_3 , a_3)

(iii) Merkitään väleihin

(F_1 , a_1) ja (a_3 , F_3)

kuuluvat havaintoarvot kuvioon *tähdillä*.

Merkitään väleihin

$(-\infty , F_1)$ ja $(F_3 , +\infty)$

kuuluvat havaintoarvot kuvioon *ympyröillä*.

Box and Whisker -kuvio 5/6

- Laatikon ja viiksien määrittelemään väliin

  (a_1 , a_3)

  kuuluvat havaintoarvoja voidaan pitää *tavallisina*.
- Erityisesti kvartiilivälin

  (Q_1 , Q_3)

  määrittelemä laatikko sulkee sisäänsä *keskimmäiset* 50 % havaintoarvoista.

Box and Whisker -kuvio 6/6

- Tähdillä ja ympyröillä merkityt, välin (a_1 , a_3) ulkopuolelle jäävät havaintoarvot ovat *poikkeuksellisia*:

  (i) Tähdillä merkityt, väleihin

  $$(F_1 , a_1) \text{ ja } (a_3 , F_3)$$

  kuuluvat havaintoarvot ovat vain *lievästi* poikkeuksellisia.

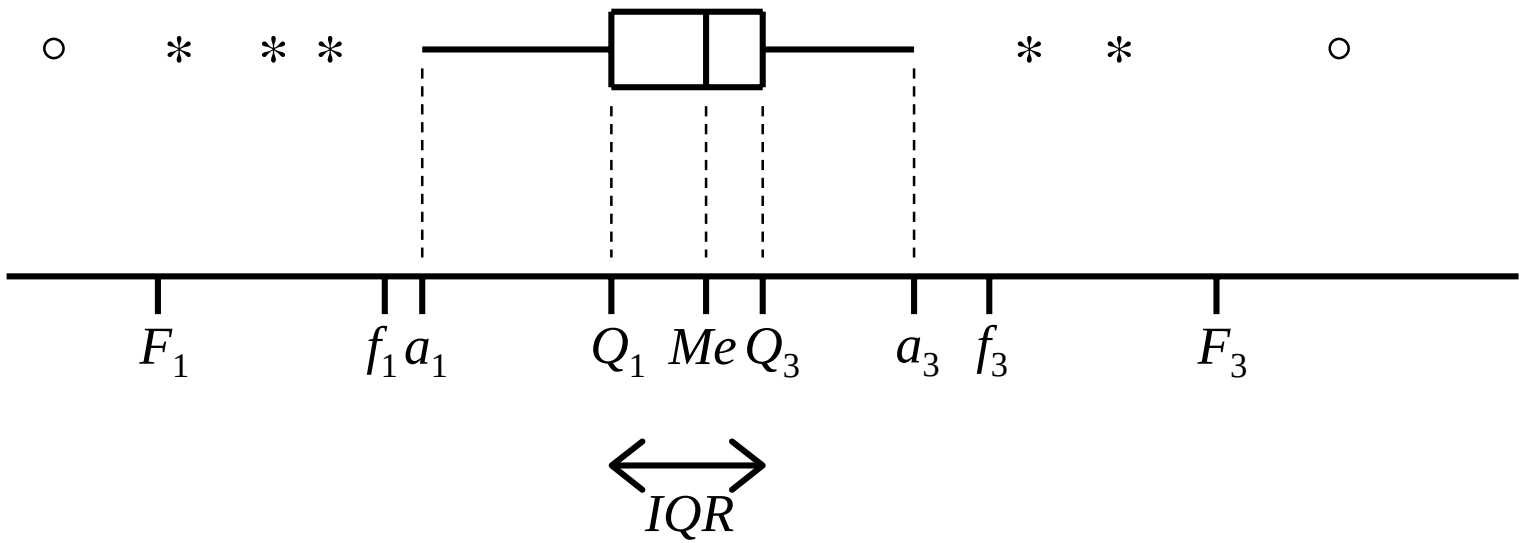
  (ii) Ympyröillä merkityt, väleihin

  $$(-\infty , F_1) \text{ ja } (F_3 , +\infty)$$

  kuuluvia havaintoarvoja voidaan pitää *voimakkaasti* poikkeuksellisina.

Box and Whisker -kuvio: Havainnollistus

- *Box and Whisker -kuvio*:



Tilastollisten aineistojen kuvaaminen

- Havaintoarvojen jakauma
- Tunnusluvut
- Suhdeasteikollisten muuttujien tunnusluvut
- Järjestysasteikollisten muuttujien tunnusluvut
- >> Laatueroasteikollisten muuttujien tunnusluvut

Laatueroasteikollisten muuttujien tunnusluvut

Avainsanat

- **Luokitellun aineiston moodi**
- **Moodi**
- **Suhteellinen frekvenssi**

Tunnusluvut laatueroasteikollisille muuttujille

- Tavallisimmat tunnusluvut *laatueroasteikollisten* muuttujien havaituille arvoille:
  - **Suhteellinen frekvenssi**
  - **Moodi** keskilukuna
- Huomautus:

  Laatueroasteikollisten muuttujien tunnuslukuja *saa käyttää ja on usein myös järkevää käyttää kuvaamaan järjestys-*, *välimatka-* ja *suhdeasteikollisten muuttujien havaittujen arvojen jakaumaa*.

Suhteellinen frekvenssi

- Olkoon *otoskoko* eli kerättyjen *havaintoarvojen lukumäärä n*.
- Olkoon *A* perusjoukon osajoukko ja otokseen kuuluvien *A*-tyyppisten havaintoarvojen *frekvenssi* eli *lukumäärä f*.
- Tällöin *A*-tyyppisten havaintoarvojen **suhteellinen frekvenssi** eli **osuus** otoksessa on

$$\frac{f}{n}$$

Moodi

- *Frekvenssijakauman* **moodi** eli *tyyppiarvo Mo* on *yleisin* havaintoarvo.
- *Luokitellun frekvenssijakauman* **moodi** eli *tyyppiarvo Mo* on siinä luokassa, jossa luokiteltua frekvenssijakaumaa *vastaava histogrammi saavuttaa maksiminsa*.
- Huomautuksia:
  - Jos käytetty luokitus *on tasavälinen*, luokitellun frekvenssi-jakauman moodi *on* siinä luokassa, jota vastaava frekvenssi on suurin.
  - Jos käytetty luokitus *ei ole tasavälinen*, luokitellun frekvenssi-jakauman moodi *ei välttämättä ole* siinä luokassa, jota vastaava frekvenssi on suurin.

Luokitellun aineiston moodi

- **Luokitellun aineiston moodi** voidaan laskea kaavalla

$$Mo = L_i + \frac{d_{i-1}}{d_{i-1} + d_{i+1}} \times c_i$$

jossa

L_i = moodiluokan alaraja

d_{i-1} = moodiluokan ja sitä alemman luokan suora-kaiteiden korkeuksien erotus

d_{i+1} = moodiluokan ja sitä ylemmän luokan suora-kaiteiden korkeuksien erotus

c_i = moodiluokan pituus

Moodi jakauman kuvaajana

- Moodi kuvaa *yleisimpien* havaintoarvojen sijoittumista havaintoarvojen jakaumassa.
- Jos havaintoarvojen jakauma *on yksihuippuinen* ja *symmetrinen*, havaintoarvojen moodi, mediaani ja aritmeettinen keskiarvo yhtyvät.
- Jos havaintoarvojen jakauma on *monihuippuinen*, jakaumalla on useita *lokaaleja moodeja*.
- Jos havaintoarvojen jakauma on *monihuippuinen*, jakauman *lokaalit moodit* antavat usein paremman kuvan jakaumasta kuin mediaani tai aritmeettinen keskiarvo.

Moodi, mediaani, aritmeettinen keskiarvo ja jakauman vinous

- Oletetaan, että *aritmeettinen keskiarvo M*, *mediaani Me* ja *moodi Mo* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen *luokitellusta frekvenssijakaumasta*.
- Jos havaintoarvojen jakauma on *yksihuippuinen*, pätee seuraava (ks. havainnollistusta seuraavalla kalvolla):

  (i) *Vasemmalle vinoilla jakaumilla*

  $$M < Me < Mo$$

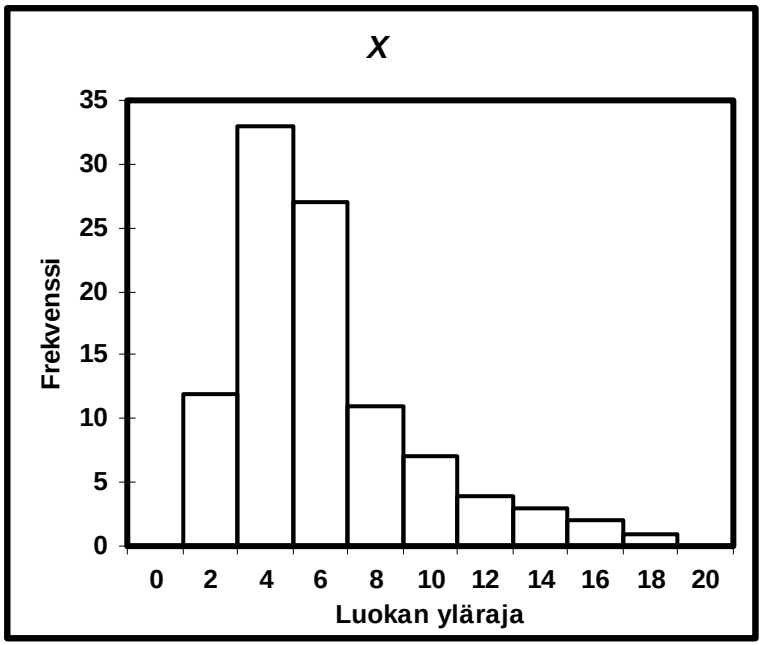
  (ii) *Symmetrisillä jakaumilla*

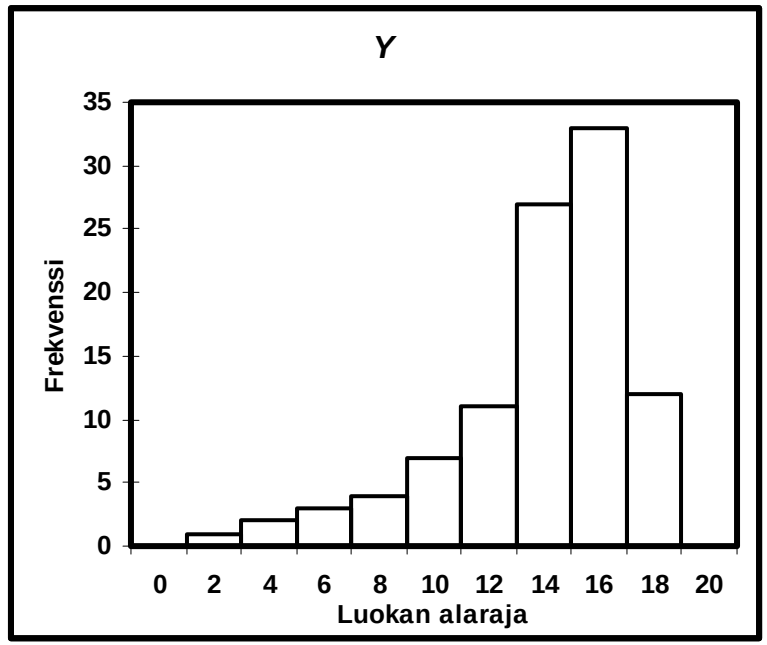
  $$M \approx Me \approx Mo$$

  (iii) *Oikealle vinoilla jakaumilla*

  $$Mo < Me < M$$

Moodi, mediaani, aritmeettinen keskiarvo ja jakauman vinous: Havainnollistus 1/2



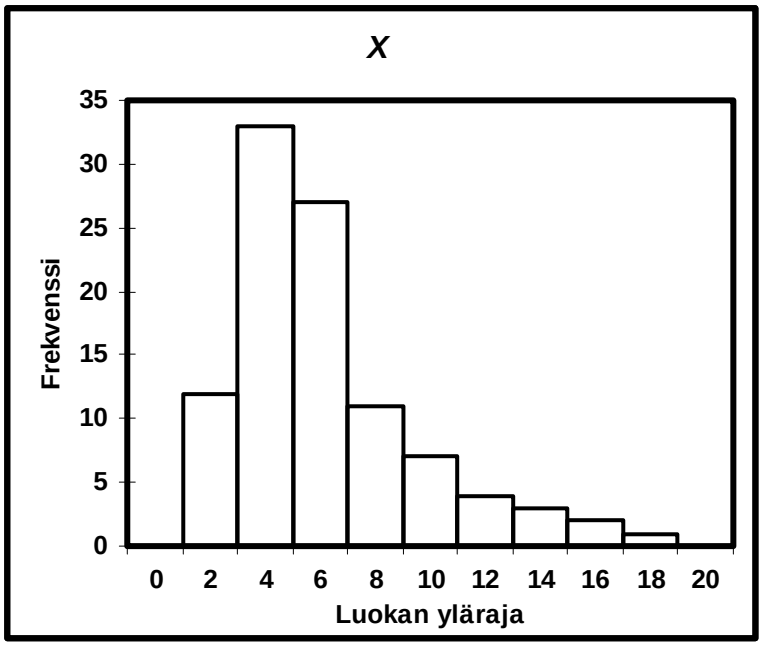


- Yllä olevat histogrammit perustuvat sataan satunnaislukujen avulla generoituun havaintoarvoon:

$$X \sim \chi^2(5)$$

$$Y = 20 - X$$

Moodi, mediaani, aritmeettinen keskiarvo ja jakauman vinous: Havainnollistus 2/2



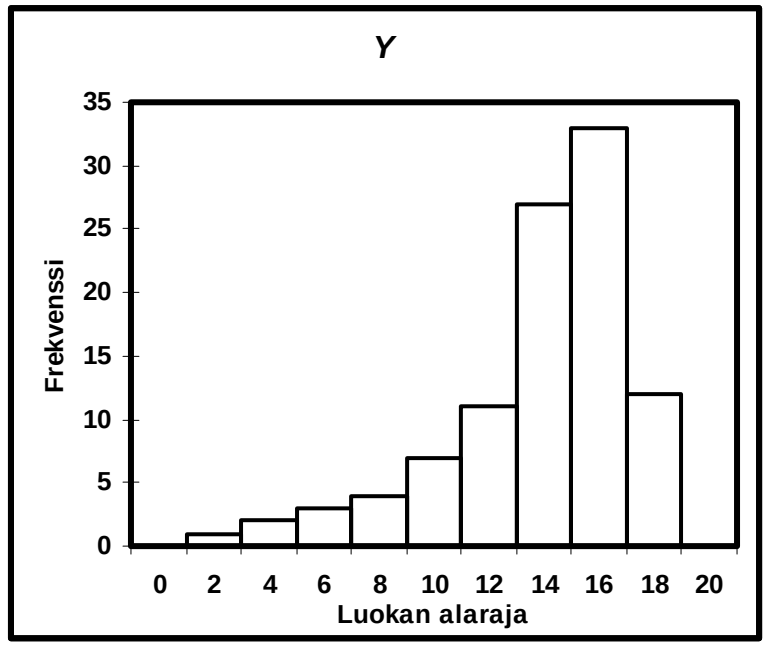
- Jakauma on vino *oikealle*:

  Vinous = 1.25

  Aritmeettinen keskiarvo = 5.19

  Mediaani = 4.41

  Moodi $\in$ (2, 4]



- Jakauma on vino *vasemmalle*:

  Vinous = –1.25

  Aritmeettinen keskiarvo = 14.81

  Mediaani = 15.59

  Moodi $\in$ (16, 18]