Hajontaluvut ja otanta

Keskiluvut eivät kerro tarpeeksi tarkasti koko havaintoaineistosta:

Esim. Kahden viikon syyslomalla mitattiin lomapäivän korkein lämpötila (asteina ℃):

I-viikko: $7,7,7,7,7,7,7$ ja II-viikko: $4,5,6,7,8,9,10$

Jolloin keskiarvo ja mediaani ovat molemmilla lomaviikoilla samat vaikka viikot olivat erityyppisiä. Tarvitaan siis lisää aineistoa kuvaavia tunnuslukuja.

Hajontaluvut ilmoittavat muuttujan vaihtelun suuruusluokan. **(vihkoon)**

Yksinkertaisin muuttujan vaihtelua ilmoittava hajontaluku on vaihteluväli (joka yleensä on kohtuu huono hajaantumisen mitta).

Määritelmä, *vaihteluväli:*

Välimatka-asteikollisen tilastomuuttujan *vaihteluväli* on muuttujan suurimman ja pienimmän arvon erotus.

Esim. I-viikko: $7-7=0$ ja II-viikko: $10-4=6$

Yritetään luoda parempi hajontaluku, joka ilmaisisi kuinka paljon havainnot eli muuttujan arvot keskimäärin poikkeavat keskiarvosta.

Määritelmä, *keskipoikkeama:*

Välimatka-asteikollisen tilastomuuttujan *keskipoikkeama* on

$$\frac{|x_1-\bar{x}|+|x_2-\bar{x}|+\cdots+|x_n-\bar{x}|}{n}=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n},$$

missä $x_i, i=1,2,3,\ldots$ ovat muuttujan arvoja (=havaintoja) ja $\bar{x}$ on keskiarvo. Miksi itseisarvo? Jotta keskipoikkeama kuvaisi keskimääräisen etäisyyden/poikkeaman…itseisarvo ⟷ etäisyystulkinta.

Esim. I-viikko: $\bar{x}=7$ ja $\frac{|7-7|+|7-7|+\cdots+|7-7|}{7}=0$

II-viikko: $\bar{x}=7$ ja $\frac{|4-7|+|5-7|+\cdots+|10-7|}{7}=\frac{3+2+1+0+1+2+3}{7}$

$$=\frac{12}{7}=1\frac{5}{7}$$

Tilastotieteen teorian kannalta parempi hajontaluku on keskihajonta, joka muodostetaan poikkeamien neliöistä ja lopuksi neliöjuuren otosta.
Syy: Näin saadaan paremmin esiin havaintojen poikkeavuus toisistaan (neliöinti kasvattaa isompia poikkeamia, $x > 1$, ja pienentää hyvin pieniä poikkeamia entisestään, $0 < x << 1$)

Määritelmä, *keskihajonta (ja varianssi):*
Välimatka-asteikollisen tilastomuuttujan *keskihajonta* on (s = standard deviation)

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}},$$

missä $x_i, i = 1,2,3, \ldots$ ovat muuttujan arvoja ja $\bar{x}$ on keskiarvo. Lisäksi usein käytetty hajontaluku *varianssi* on keskihajonnan neliö s^2.

Nämä kaksi hajontalukua ovat tärkeitä ja löytyvät laskimista, tosin voi olla eri merkinnöin: s, σ ja s^2, σ^2

Esim. I-viikko: ei mielekäs $= 0$

II-viikko: $s = \sqrt{\frac{(4-7)^2+(5-7)^2+\cdots+(10-7)^2}{n}} = 2$

Jos koko populaation sijaan tarkastellaan otosta, niin otoskeskihajonnasta (niin kuin myös keskiarvosta) saadaan *estimaatti* eli arvio koko perusjoukosta saatavalle keskihajonnalle.

Määritelmä, *otoskeskihajonta (ja varianssi):*
Olkoot $x_i, i = 1,2,3, \ldots$ havainnot. Tällöin *otoskeskihajonta* s_{n-1}on

$$s_{n-1} = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

ja *otosvarianssi* otoskeskihajonnan neliö ${s_{n-1}}^2$.

Usein n on niin iso, ettei ole väliä käyttääkö otos- vai populaatiohajontalukuja. Jos $n < 10$, niin käytä silloin ”otos”-hajontalukuja.

Keskihajonta (ja siten myös varianssi) saadaan määritettyä frekvenssien kautta, mikäli aineisto on luokiteltu ja luokkafrekvenssit tunnetaan.

Korrelaatio

Määritelmä, *korrelaatio:*
Kahden muuttujan (eli tilastoyksiköiden = tapauksien ominaisuuksien) välistä tilastollista riippuvuutta sanotaan *korrelaatioksi* (correlation = yhtäläisyys, yhteys, korrelaatio).

Esim. Miten matikan tehtävien tekeminen vaikuttaa koepisteisiin.

Sijoitetaan pisteet eli arvoparit (muuttujan 1 arvo, muuttujan 2 arvo) koordinaatistoon. Kun näin saatuun pistejoukkoon sovitetaan *regressiosuora*, niin

- kasvava suora (kulmakerroin > 0) ilmaisee pos. korrelaation ja
- vähenevä suora (kulmakerroin < 0) ilmaisee neg. Korrelaation

Regressiosuora kulkee aina keskiarvopisteen $(\bar{x}, \bar{y})$ kautta.

Keski- ja hajontaluvut kertovat yhden tilastomuuttujan arvojen jakaumasta. Kahden muuttujan tapauksessa korrelaatiota kuvaa *korrelaatiokerroin*. Lineaarisen (eli suoran omaisen) korrelaation voimakkuuden antaa *Pearsonin korrelaatiokerroin*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Tulkintaa: Aina $-1 < r < 1$ ja arvoilla $r = \pm 1$ pisteparit sijaitsevat regressiosuoralla ja korrelaatio on 100%:nen (joko neg. tai pos.). Yleisesti sanotaan, että korrelaatio on

- voimakas, jos $|r| \geq 0{,}8$
- huomattava, jos $0{,}6 \leq |r| < 0{,}8$
- kohtalainen, jos $0{,}3 \leq |r| < 0{,}8$
- merkityksetön, jos $|r| < 0{,}3$

Esim. Tarkastellaan edellisen jakson kursseja.
EXCEL- taulukkolaskentaohjelma.

Arvojen normittaminen ja vertailu

Määritelmä, *normitettu arvo:*
Muuttujan arvoa x vastaava normitettu arvo saadaan yhtälöstä

$$z = \frac{x - \overline{x}}{s},$$

jossa $\overline{x}$ on keskiarvo ja s keskihajonta.

Normitettu arvo saadaan, kun ensin siirretään keskiarvo nollaan (osoittaja) ja sitten skaalataan keskihajonnan suhteen (nimittäjä). Tämä on arkipäivää normaalijakauman tehtävissä (luku 4), toki kone hoitaa.

→ Normitettu arvo siis kertoo kuinka monen keskihajonnan päässä keskiarvosta muuttujan arvo on.
→ Normitettujen arvojen avulla voidaan verrata kumpi kahdesta eri aineistoihin kuuluvasta arvosta on poikkeuksellisempi.

Katso kirjan esimerkki 2 sivulta 39.