TILASTOT: johdantoa ja käsitteitä

TOD.NÄK JA TILASTOT, MAA10

Tilastotieteen tehtävänä on esittää ja tulkita tutkimuskohteeseen liittyvää *havaintoaineistoa* eli *tilastoaineistoa*. Tutkitaan valittua joukkoa ja sen ominaisuuksia. → Hyöty mallien ja ennusteiden luomisessa.

Määritelmä, *tilastot = havaintoaineisto:*
Tilastot ovat tutkimusta varten systemaattisesti kerättyjä ja muistiinkirjattuja havaintoja (tietoa jostakin).

Havaintojen esitystapoja ovat mm. keski-, hajonta- ja muut tunnusluvut (esimerkiksi keskiarvo), lisäksi kuviot eli diagrammit sekä taulukot ja matriisit.
Pankit, vakuutusyhtiöt, järjestöt ja toimivat elimet, yksittäiset ihmiset = sinä hyödyntävät ja käyttävät tilastoja ja todennäköisyyksiä.
Sanonta, joka kuuluu

VALHE – EMÄVALHE – TILASTOT

pitää joskus hyvin paikkansa. Mitä se tarkoittaa?

Havaintojen keruu ja muokkaus

Aluksi: Mietitään järkevä tutkimuskohde (rajaus) ja laaditaan tutkimuskysymykset joko käytännön (pakon) sanelemana tai pelkästä mielenkiinnosta.

Sitten: Datan eli aineiston keruu (tätäkin vaihetta pitää miettiä: miten, milloin, kuinka paljon/miten pitkään, millaisia virhemahdollisuuksia jne.).

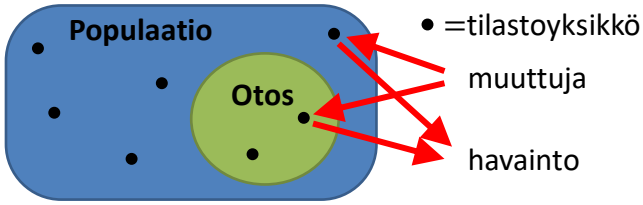
Määritelmä, *populaatio, tapaus, muuttuja:*
→ Valitaan *perusjoukko* eli *populaatio*. Perusjoukko sisältää tutkimuksen kohteet eli *tapaukset*, joita sanotaan *tilastoyksiköiksi*. Tutkittavaa (tapausten) ominaisuutta sanotaan *muuttujaksi*.

Huom! Vaarana on sekoittaa tapaukset ja muuttuja eli tapausten jokin ominaisuus, jota tutkitaan. Tapaukset mielletään x:siksi eli jonkin funktion f muuttujaksi. Nyt siis muuttuja on jokin perusjoukon tapausten ominaisuus, josta ollaan kiinnostuneita ja jota tutkitaan.

Esim. Populaatio: Sievin MAA10-ryhmän opiskelijat
Tilastoyksikkö: opiskelija
Muuttuja: opiskelijan pituus TAI
opiskelijan syntymäaika TAI
kuunteleeko opiskelija opea vai ei, jne.
Havainto: muuttujan arvo (ei tarvi olla reaaliluku)

Määritelmä, *otos:*
Usein perusjoukko on liian iso, jolloin joudutaan ottamaan *otos*, joka on sopivalla tavalla otettu, ns. *edustava*, osa perusjoukosta.

"Sopiva tapa" tarkoittaa, että jokaisella tilastoyksiköllä eli tapauksella on yhtä suuri mahdollisuus tulla valituksi otokseen.



Diskreetti ja jatkuva muuttuja

Määritelmä, *jakauma:*
Jakaumalla tarkoitetaan tilastollisessa tutkimuksessa selvitetyn havaintoaineiston muuttujan eli tutkittavan perusjoukon tapausten jonkin ominaisuuden arvoja. Millaisia arvot ovat ja miten ne ovat jakaantuneet.

Esim. Sievin lukion MAA10:n opiskelijoiden syntymäajat:
$3.1,\ 4.1,\ 26.1,\ 26.1,\ 26.1,\ 28.1,\ 4.2,\ 5.2,\ 5.2,\ \ldots, 30.12$

Määritelmä, *diskreetti ja jatkuva muuttuja sekä jakauma:*
Jos muuttuja voi saada vain *erillisiä* arvoja (esim. kurssiarvosanat: 4, $5, 6, 7, 8, 9$ ja 10, mutta ei $7{,}23$ tai $2\pi \approx 6{,}28$), niin muuttuja on *diskreetti* (*discrete*=erillinen) ja vastaava jakauma on *diskreetti jakauma*.

Jos muuttuja voi saada joltakin väliltä kaikki, eli minkä tahansa arvon, niin muuttuja on *jatkuva* (esim. pituus, ikä, kuinka kauan jaksaa kuunnella open höpinöitä) ja vastaava jakauma on *jatkuva jakauma*.

Frekvenssit ja luokittelu

Usein havaintoaineisto kootaan *taulukoiksi*. SYY: suuri informaatiotiheys pienessä koossa, tämä on positiivista. Toisaalta, pitää osata lukea taulukoita, tämä on negatiivista (monia huijataan esim. väärillä akselisuhteilla).

Määritelmä, *frekvenssi:*
Muuttujan eri arvojen esiintymiskertojen lukumäärää sanotaan *frekvenssiksi*, merkitään f. Frekvenssien yhteenlaskettu summa on sama kuin perusjoukon/otoksen tilastoyksiköiden eli tapausten lukumäärä.

Esimerkki (vihkoon) Sievin lukion MAA10 – kurssin opiskelijoiden pituudet (2 cm tarkkuudella)

$$170, 156, 172, 166, 166, 156, 162, 160, 158, 160, 160,$$

$$170, 158, 160, 166, 168, \quad \text{yht 16 kpl}$$

→ Taulukoidaan (taulu)

Määritelmä, *suhteellinen frekvenssi:*
Suhteellinen frekvenssi, joka lähes aina lasketaan ja ilmoitetaan prosentteina, on frekvenssin suhde tapausten eli tilastoyksiköiden lukumäärään. Merkitään $f\%$.

Esim. Edellisessä esimerkissä olevien pituuksien $156, 160$ ja 172 suhteelliset frekvenssit ovat:

156 cm: $\frac{2}{16} = \frac{1}{8} = 0{,}12{,}5$ eli 12,5%

160 cm: $\frac{4}{16} = \frac{1}{4} = 0{,}25$ eli 25%

172 cm: $\frac{1}{16} = 0{,}0625$ eli 6,25%

Määritelmä, *summafrekvenssit:*
Summafrekvenssi, merkitään sf, ilmoittaa kyseiseen muuttujan arvoon saakka kertyneet frekvenssit (eli lasketaan yhteen) ja *suhteellinen* summafrekvenssi, merkitään $sf\%$, vastaavasti prosenttiosuuden.

Esim. Täydennetään taulukkoa.

Jos tutkitaan samalla kertaa useita muuttujia, niin saadaan havaintomatriisi. Tämä on kustannustehokasta.

opiskelija	sukupuoli	pituus (cm)	matematiikka	
			numero	oppimäärä
Aho P	nainen	161	6	lyhyt
Airola B	nainen	156	8	pitkä
Anttila K	mies	165	9	pitkä
Asunmaa O	nainen	169	5	lyhyt
Brilli E	mies	169	7	pitkä
Cajander K	mies	178	7	lyhyt
Eskelinen S	nainen	169	9	lyhyt
Fick J	nainen	171	8	lyhyt
Halonen K	nainen	171	7	lyhyt
Huovinen V	nainen	163	9	lyhyt
Huttunen P	mies	174	10	pitkä
Korhonen L	nainen	164	8	pitkä
Leppä M	nainen	168	7	lyhyt
Sorsa M	nainen	172	7	lyhyt
Stäng J	mies	183	7	pitkä
Tenhunen P	mies	177	5	lyhyt
Vesakas F	nainen	165	7	lyhyt

Isojen tilastoaineistojen käsittely suoritetaan tietokoneilla, ohjelmistoina mm. R-ohjelmointi (ilm.) ja SPSS (maksaa/kokeiluversio ilm).

Luokittelu ja luokkakeskukset

Jatkuvan muuttujan aineisto on syytä **luokitella**. Luokittelun avulla aineiston (jossa voi siis olla paljon eri muuttujan arvoja) ominaisuudet saadaan selkeämmin esiin ja mm. tunnuslukujen laskeminen helpottuu.

Esim. Luokitellaan edellisessä esimerkissä olevat pituudet neljään luokkaan. Luokittelematon aineisto oli:

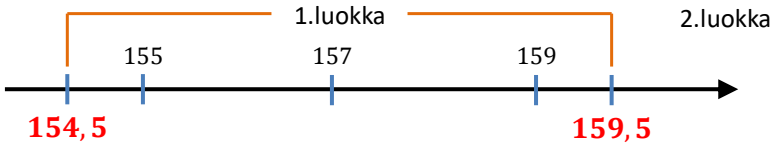
$$170, 156, 172, 166, 166, 156, 162, 160, 158, 160, 160, 170, 158, 160, 166, 168$$

Luokiteltu aineisto on (kukin muuttujan arvo *vain yhteen* luokkaan):

Luokka, cm	f	**Luokkakeskus**
$155 - 159$	4	
$160 - 164$	5	
$165 - 169$	4	
$170 - 174$	3	

Täydennetään taulukkoa. **Huom!** Luokittelu **tasa**välein, $4 - 10$ luokkaa.

Esim. (jatkuu) Luokkarajat eivät aina ole todellisia johtuen mm. pyöristyksistä. Näin ollen ensimmäisen luokan todelliset ala- ja ylärajat ovat



Luokkakeskus on luokan keskellä oleva luku. Se lasketaan ala- ja yläreunan keskiarvona

$$\text{luokkakeskus} = \frac{\text{alaraja} + \text{yläraja}}{2}$$

Täydennetään taulukkoa.

Luokka, cm	f	Luokkakeskus
$155 - 159$	4	157
$160 - 164$	5	162
$165 - 169$	4	167
$170 - 174$	3	172

Mitta-asteikot ja tunnusluvut

Havaintoaineiston muuttujat (eli tutkittavat ominaisuudet) voidaan luokitella eri mitta-asteikkojen mukaan. Asteikon valinta kertoo, millä tavoin tiettyä ominaisuutta mitataan.

Mitta-asteikot:

1. Luokittelu- eli *nominaaliasteikko*.
   Ei suuruus tai paremmuusjärjestystä. Keskiluvuista vain moodi eli tyyppiarvo soveltuu. Esimerkiksi
   sukupuoli: mies, nainen, muut (intersukupuolet)
   veriryhmä: A, B, AB, 0

2. Järjestys- eli *ordinaaliasteikko*.
   Laskutoimitukset eivät ole mielekkäitä, mutta järjestys voidaan antaa. Keskiluvuista moodi ja mediaani soveltuvat. Esimerkiksi
   sotilasarvot: alokas, jääkäri, …, kenraali

Mitta-asteikot (jatkuu):

3. Välimatka- eli *intervalliasteikko*.
   Voidaan laskea tilastoyksiköiden eli tapausten välisiä eroja = välimatkoja. Esimerkiksi lämpötila.

4. *Suhdeasteikko*.
   Voidaan laskea *muuttujien* arvojen välisiä suhteita ja verrata niitä. Esimerkiksi pituus ja massa. **Pienin arvo aina nolla!**
   Henkilö A: pituus: $170\ \text{cm}$ massa: $78\ \text{kg}$
   Henkilö B: pituus: $150\ \text{cm}$ massa: $62\ \text{kg}$
   Nyt
   $$\frac{170\ \text{cm}}{150\ \text{cm}} = 1{,}1\overline{3} \qquad \text{ja} \qquad \frac{78\ \text{kg}}{62\ \text{kg}} = 1{,}25$$
   henkilö B:n massa on siis suhteessa pituuteen pienempi kuin henkilö A:n. (Onko A lihavampi? → ei välttämättä → bodari). Eli ei pidä tehdä vääriä johtopäätöksiä! Jotta olisi suhteessa sama massa, niin verranto antaa
   $$\frac{170}{150} = \frac{78}{x} \quad \Longrightarrow x = \frac{78 \cdot 150}{170} \approx 68{,}823\ldots\ \text{kg}$$

Keskiluvut

Keskiluvuilla kuvataan havaintoarvojen keskimääräistä suuruutta tai laatua.

Keskiluvut:

1. *Keskiarvo*.
   Ilmoittaa muuttujan keskimääräisen suuruuden. On tunnetuin keskiluku, merkitään $\bar{x}$ (ei vektorimerkintä), jonka yhtälö on
   $$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$
   Keskiarvo voidaan laskea myös frekvenssien kautta, ns. painotettuna keskiarvona. Kun n on kaikkien havaintojen lukumäärä ja k on eri havaintojen, eli luokkien lukumäärä, niin
   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i\,, \qquad k \leq n$$

Keskiluvut (jatkuu):

2. *Moodi eli tyyppiarvo*.
   Moodi, lyhennetään ja merkitään (Mo), on suurinta frekvenssiä eli suurinta muuttujan arvon esiintymiskerran lukumäärää, vastaava lukuarvo. (Kuinka monta kertaa esiintyy eniten esiintyvä muuttujan arvo.) Moodi siis soveltuu kaikille mitta-asteikoille.

3. *Mediaani*.
   Mediaani, lyhennetään ja merkitään (Md), on järjestykseen asetetun havaintoaineiston keskimmäinen arvo (pariton määrä) tai kahden keskimmäisen arvon keskiarvo (parillinen määrä). Mediaani vaatii järjestyksen eli se soveltuu muille paitsi luokitteluasteikolle, jossa ei ole järjestystä. Muistisääntö: Kun aineisto järjestyksessä, niin koko havaintoaineistosta on 50% mediaanin molemmilla puolilla.

Esim. Jonkin muuttujan havaintoarvot ovat

$1, 3, 2, 5, 2, 2, 3, 8, 2, 2, 5, 5, 3, 2, 4, 8, 4, 278,$ yht 18 kpl.

Mo

→ havaintoarvo: 1 **2** 3 4 5 6 7 8 278

frekvenssit: 1 6 3 2 3 0 0 2 1 $\sum$ =18

Keskiarvo: $\bar{x} = \frac{1+3+2+\cdots+278}{18} = \frac{1\cdot1+6\cdot2+3\cdot3+\cdots+1\cdot278}{18} = 18{,}8\overline{3}.$

Moodi: $Mo = 2$, koska kakkonen esiintyy 6 kertaa.

Mediaani: Laitetaan ensin havaintoarvot järjestykseen!

$1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 8, 8, 278$

9kpl eli 50% 9kpl eli 50%

Nyt mediaani on kahden keskimmäisen luvun keskiarvo, eli $Md = (3 + 3)/2 = 3$.

Keskiluvut (jatkuu):

1. *Painotettu keskiarvo*.
   Lukujen $x_1 + x_2 + \cdots + x_n$ painotettu keskiarvo, kun painoina ovat luvut $p_1 + p_2 + \cdots + p_n$ (> 0), on

$$\frac{p_1 x_1 + p_2 x_2 + \cdots + p_n x_n}{p_1 + p_2 + \cdots + p_n}$$

Esim. Eräässä matematiikan kokeessa arvosanojen jakauma oli seuraava:

4	5	6	7	8	9	10
1,3 %	9,8 %	15,8 %	20,3 %	23,3 %	23,4 %	6,1 %

Laske arvosanojen keskiarvo.

Ratkaisu Nyt prosentit ovat ns. painoja, jotka summautuvat ykköseksi, saadaan

$$\overline{x} = \frac{0{,}013 \cdot 4 + 0{,}098 \cdot 5 + \cdots + 0{,}061 \cdot 10}{1} = 7{,}491 \approx 7{,}5\,.$$