

## KYSYMYKSIÄ JA ESIMERKKEJÄ TILASTOTIETEESTÄ JA TILASTOJEN LUKUTAIDOSTA

**Tekijät:** Saara Sarsa, Eeli Tamminen, Maiju Mäenpää, Amanda Sirén ja Katri Pietiläinen

**Laadittu** keväällä 2023.

Tämä teos on lisensoitu Creative Commons Nimeä 4.0 Kansainvälinen -lisenssillä.

**Kohderyhmä:** yläkoulu, lukio

**Esitiedot:** tilastotieteen peruskäsitteet (keskiarvo, mediaani, moodi), otos, frekvenssi, jakaumat, aineiston havainnollistaminen visuaalisesti, kohdassa 5 vektorit

**Oppimistavoitteet:** Esimerkin/kysymyksen käsittelyn jälkeen oppilas...

- osaa arvioida tilastoja ja tulosten yleistettävyyttä.
- ymmärtää tilastojen tulkintaan liittyviä virhetilanteita.
- vahvistaa ymmärrystään tilanteeseen sopivan tilastollisen mallin merkityksestä.

Materiaalissa on käytetty lähteenä Tilastokeskuksen ”Opi tilastoista” -sivustoa (luettu 12.2.2023) ja opasta ”Kuinka teen tilastotutkimuksen?” (luettu 12.2.2023).

### SISÄLLYS

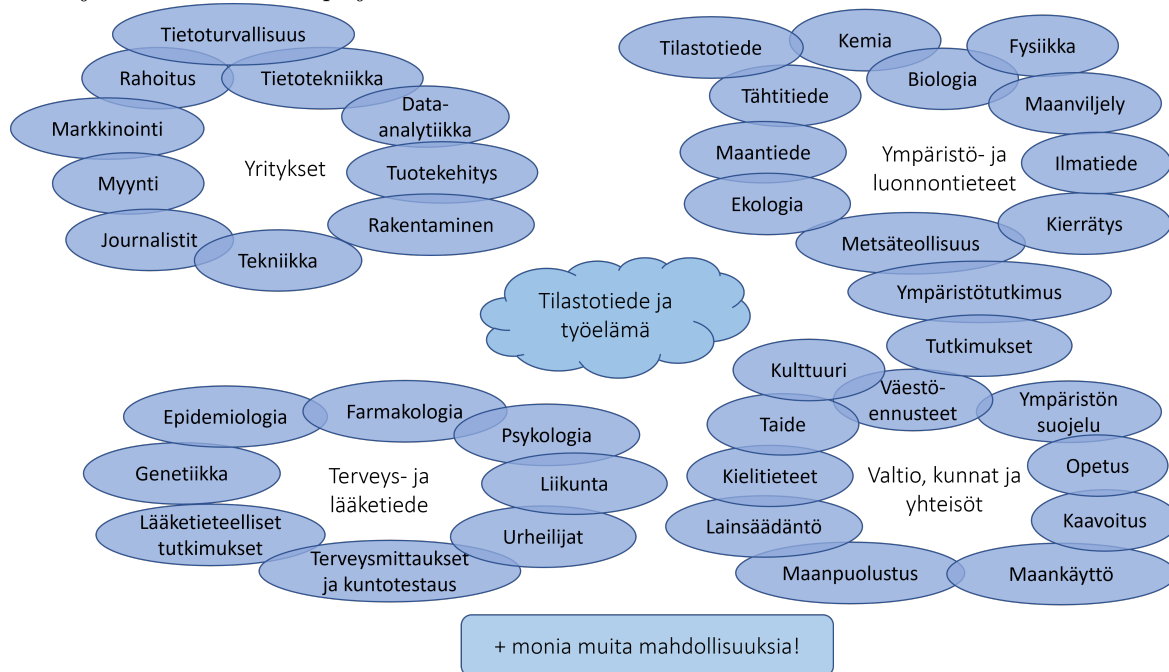
1. Mihin tilastoja tarvitaan?	2
2. Missä ammateissa tarvitaan tilastoja?	2
3. Miksi tilastolukutaitoa tarvitaan?	2
4. Miten tilastollista tutkimusta tehdään?	2
5. Miten aineistoja analysoidaan tilastollisesti?	3
6. Mikä on pienimmän neliösumman menetelmä?	3
7. Mistä dataa saadaan?	3
8. Simulaatio: Sadan nopan heitto.	3

## 1. MIHIN TILASTOJA TARVITAAN?

**Vastaus:** Tilastojen avulla saamme tietoa maailmasta ja ne koostuvat tiedosta, jota on saatu mittausten avulla. Ne tukevat teoreettiseen tietoon ja faktaan perustuvaa päätöksentekoa.

## 2. MISSÄ AMMATEISSA TARVITAAN TILASTOJA?

**Vastaus:** Tilastojen käsittelyä ja ymmärrystä tarvitaan lähes kaikissa ammateissa. Alla olevassa kuvassa esitellään esimerkkiammattia ja -aloja, joissa tarvitaan tilastotieteellistä ymmärrystä. Todellisuudessa näitä aloja on olemassa vielä paljon enemmän.



## 3. MIKSI TILASTOLUKUTAITOA TARVITAAN?

**Vastaus:** Vastataan kahden palikkaesimerkin kautta.

- (1) Vuonna 2017 suomalaisten miesten keskimääräinen pituus oli 177,5 cm ja naisten 163,2 cm. (Lähde: Terveyden ja hyvinvoinnin laitoksen (THL) FinTerveys-väestötutkimus, luettu 12.2.2023.)
  - Tarkoittaako tämä, että *kaikki* miehet ovat pidempiä kuin *kaikkia* naiset? (Ei.)
  - Mitä tiedetään pituuskokouksesta tämän perusteella? (Ei mitään muuta kuin mihin keskiarvo sijoittuu.)
  - Ovatko nämä luvut luotettavia? (Riippuu otoskokoista ja otoksen edustavuudesta. Ovatko linjassa arkihavaintojen kanssa?)Johtopäätös: Pelkkä keskiarvo ei vielä anna kattavaa kuvaa ilmiöstä. Tarvitaan muita tilastollisia menetelmiä ja tutkimusasetelman arviointia.
- (2) Korrelaatio vs. kausaalisuus. Nämä menevät usein sekaisin, esim. mediassa.
  - Korrelaatio: Jäätelönsyönnin ja hukkuminen ovat tilastollisesti yhteydessä toisiinsa.
  - Kausaalisuus: Jäätelönsyönnin aiheuttaa hukkumista?Muita samanlaisia esimerkkejä Googelta hakusanalla **spurious correlations**.

## 4. MITEN TILASTOLLISTA TUTKIMUSTA TEHDÄÄN?

**Vastaus:** Tilastollisen tutkimuksen lähtökohta on tutkimusongelma tai hypoteesi, jolle haetaan vahvistusta havainnoimalla ja keräämällä tietoa tutkittavasta ilmiöstä ja kokoamalla niistä tilasto/dataa. Tietoja tutkimukseen voidaan kerätä mittauksin, kyselyillä tai saada rekistereistä. Läheskään aina ei pystytä keräämään tietoa kaikista tutkimukseen kohteena olevista populaation jäsenistä, vaan päädytään tutkimaan vain osaa populaatiosta.

Kerätty aineisto tallennetaan muotoon, jota voidaan käsitellä tietokoneilla ja sen pohjalta voidaan laskea tilastollisia tunnuslukuja, laatia kuvaajia sekä muotoilla matemaattisia/tilastollisia malleja. Tämä vaihe tehdään tietokoneohjelmistojen avulla.

Mallin sopivuus ja tulosten perusteella tehtävät johtopäätökset perustuvat tilastollisen päättelyyn, joka puolestaan perustuu todennäköisyyslaskentaan. Tässä vaiheessa arvioidaan ovatko tulokset tilastollisesti merkitseviä eli voisivatko on olla sattuman tulosta vai ei.

Käytetyn mallin, siihen liittyvien mallioletusten ja menetelmien sopivuus täytyy ottaa huomioon koko tutkimusprosessin ajan, jotta tuloksen avulla tehtävät päätelmät olisivat luotettavia.

#### 5. MITEN AINEISTOJA ANALYSOIDAAN TILASTOLLISESTI?

**Vastaus:** Pieniä aineistoja analysoidaan kynällä ja paperilla. Suurempia, esimerkiksi 10 000 havainnon, aineistoja analysoidaan ohjelmallisesti. Ohjelmistoilla tehtävä analyysi pohjautuu aineiston tarkasteluun vektoreina ja matriiseina (taulukoina). Näistä vektoreista ja matriiseista voidaan määrittää tilastollisia tunnuslukuja vastaavilla laskukaavoilla.

Suurten aineistojen analysointiin voi käyttää monia eri ohjelmistoja. Vaihtoehtoina ovat esimerkiksi taulukkolaskentaohjelmat (Microsoft Excel, LibreOffice Calc), erilaiset ohjelmointikielet (Python, R, SAS), GeoGebra ja SPSS. Edistyneemmillä ohjelmilla voidaan analysoida ja muodostaa helposti erilaisia malleja.

#### 6. MIKÄ ON PIENIMMÄN NELIÖSUMMAN MENETELMÄ?

**Vastaus:** Pienimmän neliösumman menetelmällä pyritään löytämään lineaariselle regressiomallille sovi-tesuora eli malli, joka kuvaa selitettävän muuttujan vaihtelua. Menetelmässä regressiokertoimien arvot määrätään minimoimalla jäännös- eli virhetermien neliösumma.

Virhetermi tarkoittaa tässä sovitesuoran ja havainnon välistä pystysuoraa etäisyyttä.

#### 7. MISTÄ DATAA SAADAAN?

**Vastaus:**

- oma tutkimus
- simulaatio eli usein tietokoneella toteutettu todellisuuden jäljittely, ks. kohta 8
- avoin data

#### 8. SIMULAATIO: SADAN NOPAN HEITTO.

**Vastaus:** Simulaatio on kehitetty Abitin Python3-editorin kirjastoilla.

```
import random
import statistics

# 100 heiton simulaatio

lukulista = [] # lista heittojen tallennukseen

# heitetään noppaa 100 kertaa
for i in range(0,100):
    lukulista.append(random.randint(1,6))

# aineiston tunnuslukuja
print(statistics.mean(lukulista)) # keskiarvo
print(statistics.mode(lukulista)) # moodi
print(statistics.stdev(lukulista)) # keskihajonta
```