

TILM 3701 - Tilastotiede ja Data 2022

Koonneet Henri Nyberg¹ Roope Rihtamo²

2022-08-22

¹Turun Yliopisto, matematiikan ja tilastotieteen laitos, henri.nyberg@utu.fi

²Turun Yliopisto, matematiikan ja tilastotieteen laitos, roope.rihtamo@utu.fi

Contents

Tilastotiede ja kurssin idea	5
Tilastotieteen asema tutkimusyhteisön ulkopuolella	7
Kurssin luonne tilastotieteen (ja datatieteen/data-analytiikan) opin- tojen esittelijänä	7
1 Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa	9
1.1 Tieteellinen ajattelu tietoyhteiskunnan perustana	10
1.2 Tilastojen yleisestä roolista yhteiskunnassa	10
1.3 Mitä on tiede?	10
1.4 Mitä on tutkimus?	11
1.5 Tieteellisen menetelmän kriteereitä	11
1.6 Tieteellinen tutkimuksen vaiheet ja tulosten julkaiseminen	11
2 Tilastotiede tieteenalana	13
3 Sattuma ja satunnaisuus	15
4 Tilastolliset aineistot, niiden kerääminen ja mittaaminen	17
5 Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma	19
6 Tilastollinen riippuvuus ja korrelaatio	21
7 Regressioanalyysi	23
8 Tilastotieteen rooli uuden tiedon tuottamisessa	25

9	Aineisto- ja tutkimustyytit ja koeasetelmat	27
10	Tilastollisesta ennustamisesta	29
11	Tilastotieteen kehityksen nykytrendejä	31

Tilastotiede ja kurssin idea

- Tämän tilastotieteen ensimmäisen kurssin ideana on (ainakin)
 - Esitellä ja johdatella **tilastolliseen ja tieteelliseen ajatteluun** ja sen hyödyntämiseen eri tyyppisissä tutkimusongelmissa.
 - Esitellä tilastotieteen roolia **empiirisen tutkimusaineiston keräämisessä ja analyysissä** sekä tarkastella tieteentekemisen ja tilastotieteen suhdetta.
 - Pohtia **tilastotieteen olemusta tieteenalana** ja tarkastella tilastotieteen ja datatieteiden (data sciencen) samankaltaisuuksia ja eroja.
 - Pohtia **sattuman ja satunnaisuuden roolia** jokapäiväisessä elämässä ja erityisesti osana tieteellistä tutkimusprosessia.
 - Oppia tilastotieteen peruskäsitteitä ja (tilastollisen) tutkimuksenteon alkeita ja siihen liittyviä mahdollisia ongelmia esimerkiksi tilastollisten aineistojen keräämisessä.
 - Oppia tilastollisten aineistojen **kuvaamisen ja käsittelyn** alkeita sekä tilasto(tieteellisen)llisen **mallintamisen** ja **koeasetelmien** peruskäsitteitä.
- Kurssilla käsitellään myös **tilastollisen päättelyn** peruskäsitteitä ja perusteita kuten
 - Mitä on **todennäköisyys** ja miten sen tulkitaan tilastotieteessä sekä laajemmin tieteessä. Erityisesti tilastotieteen osalta keskiössä on tämän kurssin osalta **satunnaismuuttujat** sekä niihin liitettävät käsitteet
 - * **Odotusarvo, varianssi** ja kahden (tai useamman) satunnaismuuttujan **korrelaatio**.
 - * Satunnaismuuttujien **todennäköisyysjakaumien** perusteita ja niiden yhteyksiä mm. normaalijakaumaan ja muutamiin muihin keskeisiin jakaumiin.

- * Tilastollinen malli työkaluna satunnaismuuttujien formaalissa mallintamisessa ja päättelyssä. Tilastollisen malliin liittyy (usein) **parametreja** joihin tilastollinen päättely kohdistuu.
 - * Tilastollisten mallien **estimoinnin** perusidea, eli miten tilastollisen mallin parametreille muodostetaan arvot käytettävissä olevan aineiston pohjalta. Esimerkiksi: mitä tarkoittaa tilastollisen mallin parametrin **estimaattori** ja sen **harhattomuus**?
 - * Alustavia tarkasteluja tilastollisen mallin uskottavuuden käsitteelle ja **luottamusväleille** tilastollisen mallin estimoiduille parametreille.
-
- Toinen kurssin keskeisistä teemoista on tarkastella tieteellistä tutkimusprosessia teoriassa ja käytännössä. Tämä sisältää mm. seuraavia aiheita (joita siis käsitellään tällä kurssilla päällisin puolin ja varsin yleisestä näkökulmasta katsoen): tarkemmat yksityiskohdat jäävät tätä kurssia seuraavien tilastotieteen kurssien aihepiireiksi):
 - **Tutkimusongelman** asettaminen: mitä halutaan tutkia?
 - Tutkimusongelman täsmentäminen ja **tutkimusstrategian** laatiminen: millä keinoin asetettuun tutkimusongelmaan voidaan vastata?
 - **Tutkimusaineiston** (tai vain lyhyemmin **aineiston** eli **datan**) kerääminen
 - * **Aineiston ennakkoehdot**: mitkä ehdot tulee täyttyä, jotta asetettuun tutkimusongelmaan voidaan vastata?
 - * **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää aineiston ennakkoehdot? Erilaisissa tutkimuksissa käytetään erilaisia aineistoja kuten:
 - Survey- ja rekisteriaineistot
 - Havaintoarvojen välistä korrelaatiota esiintyy mm. aikasarja-aineistojen tai pitkittäisaineistojen tapauksessa
 - **Aineiston kuvaaminen**: minkälaista aineistoa on kerätty ja vastaako se ennakkoehtoja?
 - **Aineiston analyysin** lähtökohtia
 - Mitä tilastollista mallia/malleja käytetään?
 - Mitä tarkoitetaan mallien tuntemattomien parametrien arvojen estimoinnilla?
 - Tilastollinen päättely (estimointitulosten pohjalta)
 - **Johtopäätelmien** tekeminen tilastollisen päättelyn pohjalta: saatiinko tutkimusongelmaan vastaus ja kuinka luotettava saatu vastaus on?

Tilastotieteen asema tutkimusyhteisön ulkopuolella

- Tilastotiede on oppiaineena usein varsin tuntematon toisen asteen opinnoista valmistuneelle, sillä sitä ei juurikaan opeteta lukioissa tai ammattikouluissa huolimatta sen keskeisestä ja kasvavasta roolista tiedemaailman kentillä.
- Tiedeyhteisön ulkopuolellakin **tilastotiedettä ja tilastotieteilijöitä arvostetaan laajalti**.
- **Tilastotiede onkin nostanut profiliaan viimeisten vuosikymmenien aikana** tietoteknisen kehityksen tuotua laajat tietoaineistot ja kehittyneet laskennalliset menetelmät lähes jokaisen kansalaisen saataville.
- Tämä “datavallankumous” näkyy tilastotieteilijöiden kysynnässä työmarkkinoilla: erilaisten aineistojen määrän lisääntyessä kasvaa myös kysyntä työntekijöistä, jotka osaavat ammatitaitoisesti käsitellä, tulkita ja mallintaa tilastollisia aineistoja.
- Ei siis liene ihmeäkään, että erilaisten “data”-alkuisten työpaikkojen, kuten **datatieteilijä** (eng. **data scientist**) tai **data-analyytikko** (**data-analyst**) määrä on kasvanut voimakkaasti jo pidempään. Kaikkia tieto- ja datainensiivisten ammattien tekijöitä yhdistää yksi tekijä: **heidän tulee hallita ja osata tilastotiedettä!** Karkeistettuna mitä paremmin ja enemmän (laajemmin), sen parempi palkka ja monipuolisemmat työtehtävät!

Kurssin luonne tilastotieteen (ja datatieteen/data-analytiikan) opintojen esittelijänä

Kurssin mittaan esitellään tilastotieteen perusteiden lisäksi **miten TY:ssa tilastotieteen opinnoissa syvennyttään** tällä kurssilla esiteltäviin menetelmiin, aineistotyyppeihin ja mallinnuskokonaisuuksiin.

Chapter 1

Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa

Tässä luvussa tarkastellaan tieteen ja tieteellisen tutkimusprosessin luonnetta erityisesti uuden **tutkitun** tiedon tuottamisen näkökulmasta. Tiedelukutaidon merkitys on kasvanut nyky-yhteiskunnassa, kun tiedejulkaisujen saavutettavuus ja tunnettuus on lisääntynyt mm. tieteen popularisoinnin ja median laajemman tiedeuutisoinnin vuoksi. Voidakseen ymmärtää ja arvioida kriittisesti tiedeuutisia tulee lukijan olla tietoinen tieteellisen tutkimuksen luonteesta: miten tutkimusartikkeleja luetaan, mitä niiltä voidaan odottaa ja minkälaiset tulokset ovat uskottavia. **Tilastotiede näyttelee keskeistä roolia lähes kaikessa tutkimuksessa ja erityisesti erilaisten tutkimuskysymysten ja niitä vastaavien hypoteesien testauksessa.** Aloitetaankin kurssin opimateriaalin käsittely määrittelemällä ensimmäinen tilastotieteen perustermi: hypoteesi.



Hypoteesi

- Hypoteesi tarkoittaa (tausta)teorioista johdettua tai aikaisemman tutkimuksen perusteella esitettyä ennakoitua ratkaisua tai selitystä tutkittavaan ongelmaan.
- Hypoteesi ilmaistaan väitteenä, jonka paikkansapitävyyttä halutaan tutkia
- Kokeelliset tiedot voivat osoittaa hypoteesin vääräksi
- Nollahypoteesi vastaa tavallisesti tyypillistä, odotettavissa olevaa tulosta, esimerkiksi ettei kahden mitatun ilmiön välillä ole yhteyttä tai että tietty hoito on tehotonta
- Nollahypoteesia ei todisteta (“hyväksytä”), vaan voidaan ainooastaan sanoa, ettei aineisto tarjoa todistusaineistoa (“evidenssiä”) nollahypoteesin hylkäämiselle – ts. sille tulemalle, että emme hylkää nollahypoteesia.
- Vastahypoteesi sisältää usein mielenkiinnon kohteena olevan tapahtuman, kuten “on eroa” tai “on vaikutusta”
- Tutkijoilla on usein taipumus jättää julkaisematta tutkimustuloksia, joissa nollahypoteesi jää voimaan. Yleensä tämä tilanne syntyy, kun lopputulos ei eroa jo aikaisemmin otaksutusta. (Toki ajoittain tilanne on myös toisinpäin)

Tähän joku esimerkki vielä?

1.1 Tieteellinen ajattelu tietoyhteiskunnan perustana

Kesken vielä.

1.2 Tilastojen yleisestä roolista yhteiskunnassa

Kesken vielä.

1.3 Mitä on tiede?

Kesken vielä.

1.4 Mitä on tutkimus?

Kesken vielä.

1.5 Tieteellisen menetelmän kriteereitä

Kesken vielä.

1.6 Tieteellinen tutkimuksen vaiheet ja tulosten julkaiseminen

Tieteellinen tutkimus ja asiantuntijatyö tuottavat valtavan määrän perusteltua, luotettavaa tutkimustietoa. Ks. tarkemmin tieteellisestä julkaisemisesta linkin tapauksessa erityisesti yhteiskuntatieteiden alalla, mutta peruseriaatteet pätevät myös muiden tieteenalojen tapauksessa

<https://blogs.uef.fi/tiedonhaku-yhteiskuntatiede/tieteelliset-julkaisut/>

Vastuullisen tieteen

<https://vastuullinentiede.fi/fi/julkaiseminen>

artikkelit tarjoavat tietoa siitä, kuinka tutkittua tietoa tuotetaan, julkaistaan ja arvioidaan luotettavasti ja yhteisesti hyväksytyllä tavalla. Jotta tiede vaikuttaa koko yhteiskunnan hyväksi, toiminnan on oltava vastuullista tutkimuksen jokaisessa vaiheessa.

- Julkisuus ja avoimuus tekevät tutkimuksesta tiedettä.
- Tiedeyhteisöjen sisäistä ja ulkoista tiedonvälitystä ja vuorovaikutusta. Tutkimuksesta viestiminen ei ole vain tutkimustuloksista viestimistä. Vastuullinen tiedeyhteisö lisää luottamusta tieteelliseen tietoon.
- Tieteellinen julkaiseminen on tutkijoille tärkeä meritoitumisen tapa, ja siksi on tärkeää, että tekijäyhteisö määrittää niin, että se palkitsee tutkijat oikeudenmukaisesti.

Chapter 2

Tilastotiede tieteenalana

Chapter 3

Sattuma ja satunnaisuus

Chapter 4

Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Chapter 5

Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma

Chapter 6

Tilastollinen riippuvuus ja korrelaatio

Chapter 7

Regressionanalysis

Chapter 8

Tilastotieteen rooli uuden tiedon tuottamisessa

Chapter 9

Aineisto- ja tutkimustyyppit ja koeasetelmat

Chapter 10

Tilastollisesta ennustamisesta

Chapter 11

Tilastotieteen kehityksen nykytrendejä