Luku 5 - Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Tiivistelmä

Luvun ydinviesti

Tässä luvussa tarkastellaan sitä, miten tilastollisen tutkimuksen keskeinen työaskel eli otanta tehdään.

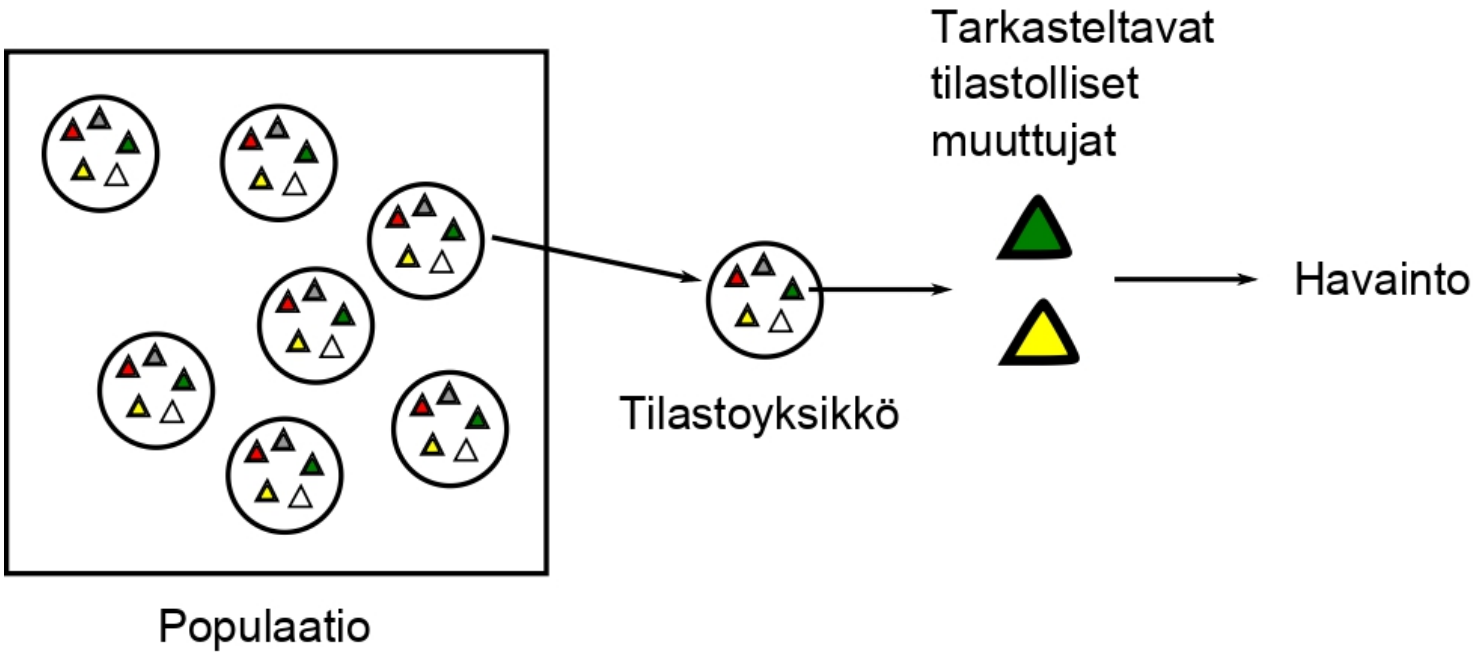
Muistetaan että tilastollisen tutkimuksen tarkoituksena on pyrkiä tekemään populaatiota koskevia yleistyksiä sitä kuvaavan havaintoaineiston perusteella.

Otannan tarkoitus on poimia tai hankkia aineisto niin että se kuvaa tutkimuksen kohteena olevaa populaatiota **edustavasti**.

Erilaisten tutkimuskysymysten tutkiminen määrittää sen populaation, johon yleistyksiä halutaan tehdä, joten se määrää myös kohteena olevan **perusjoukon**, josta havaintoaineisto kerätään, sekä käytettävän otantamenetelmän.

Tämä valinta ei ole kuitenkaan aivan yksinkertaista!

Kertausta: data eli aineisto



Tilastollinen tutkimusaineisto, eli havaintoaineisto, koostuu tilastoyksiköiden muodostamasta populaatioista esimerkiksi otannalla poimituista alkioista.

Näiltä tilastoyksiköiltä havaitaan tai **mitataan** tutkimuksessa tarkasteltavat tilastolliset muuttujat.

Havaintoaineisto

	tilastomuuttuja 1	tilastomuuttuja 2	...	tilastomuuttuja m
tilastoyksikkö 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$
tilastoyksikkö 2	$x_{2,2}$	$x_{2,2}$	...	$x_{2,m}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
tilastoyksikkö n	$x_{n,1}$	$x_{1,2}$	...	$x_{n,m}$

Usein havaintoaineisto voidaan koota esimerkiksi yllä olevan kuvan mukaiseksi taulukoksi.

Tilastoyksiköt ladotaan riveille allekkain ja näihin liitettävät tilastomuuttujat asetetaan sarakkeisiin. Yo. taulukossa on siis n tilastoyksikköä, joista jokaisesta on kerätty m tilastomuuttujan arvot.

Tilastollisen muuttujan i tilastollisesta muuttujasta j havaittu arvoa merkitään esimerkiksi $x_{i,j}$. Varsinaisen tutkimuksen kannalta mielenkiintoisia muuttujia kutsutaan **tutkimusmuuttujiksi** ja muita **taustamuuttujiksi**.

Keskeiset termit: kokonaistutkimus ja otantatutkimus

Kokonaistutkimus

- Kokonaistutkimus on tutkimus, jossa tutkitaan kaikki tutkimuksen kohteena olevan perusjoukon alkiot, ts. kaikki ajateltavissa olevat kohteet tutkitaan eli havaintoaineisto koostuu koko populaatiosta.

Otantatutkimus

- Otantatutkimuksessa tutkimus kohdistetaan johonkin populaation, eli perusjoukon, osajoukkoon, joka poimitaan sopivaa **otantamenetelmää** käyttäen. Tähän osajoukkoon poimittuja alkioita kutsutaan **otosyksiköiksi** ja koko otosjoukkoa **otokseksi**.
- Tästä osajoukosta tehdään päätelmiä, jotka yleistetään koskemaan koko populaatiota/perusjoukkoa.

Otannan idea

Otantatutkimuksen (karkeat) suunnittelu- ja työvaiheet sisältävät mm. tutkimuksen tavoitteiden asettamisen, näitä vastaavan perusjoukon asettamisen, otannan suorittamisen sekä lopulta aineiston analysoinnin ja tulosten raportoimisen.

Tavoitteena on poimia **edustava otos** mielenkiinnon kohteena olevasta perusjoukosta, sillä tämä mahdollistaa otoksesta tehtyjen päätelmien yleistämisen koko perusjoukolle!

Edustavuus
Tutkimukseen valittu perusjoukon osajoukko, otos, kuvaa perusjoukon ominaisuuksia kattavasti.

Mittaaminen

Kvantitatiivisen tutkimuksen aineistoksi kelpaa kaikki havaintoihin perustuva informaatio, joka on **mittauksen** avulla muutettavissa numeeriseen muotoon. Havaintoyksiköiden tilastollisten muuttujien numeerisia arvoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi.**

Tilastollisessa tutkimuksessa tilastomuuttujat ovat satunnaisia ja tavoite on mittaamalla liittää jokin luku satunnaisilmiötä kuvaavaan ominaisuuteen, eli mitata satunnaismuuttujan havaittua arvoa.

Hyvä mittari on **(i) validi**, eli mittaus esittää mitattavaa ominaisuutta oikein ja **(ii) luotettava**, eli mittaus on **harhaton** ja **toistettavissa**.

Kun käytetään hyviä mittareita, voidaan luotettavuutta vielä erikseen tarkastella laskemalla aineistosta tunnuslukujua mittauksen luotettavuudelle, esimerkkinä **luottamusväli**.

Keskeiset termit: harhattomuus ja toistettavuus

Harhattomuus
Mittari on harhaton, jos se ei systemaattisesti ali- tai yliarvioi mitattavan ominaisuuden määrää.

Toistettavuus
Mittari on toistettava, jos se tuottaa keskimäärin samanlaisia mittauksia samanlaisista otoksista eli se on johdonmukainen ja mittausvirheet pieniä.

Kun mittaaminen on luotettavaa ja validia, tutkimusaineisto on **sisäisesti luotettavaa**.

Aineisto on **ulkoisesti luotettavaa** silloin, kun tutkittu otos edustaa perusjoukkoa eli on edustava.

Mitta-asteikot

Laatueroasteikko/luokitteluasteikko: muuttujan mittaustaso on sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin, mutta luokkien järjestyksellä ei ole merkitystä. Esim: veriryhmä tai kotikunta. Kvalitatiivinen.

Järjestysasteikko (ordinaaliasteikko): muuttujan arvot voidaan luokittelun lisäksi asettaa empiirisesti mielekkääseen järjestykseen mitattavan ominaisuuden perusteella. Esim sotilasarvo, syntymäkuukausi. Kvalitatiivinen.

Välimatka-asteikko (intervalliasteikko): luokittamisen ja järjestyksen asettamisen lisäksi havaintoarvojen välimatkalla on empiirisesti mielekäs tulkinta, ts. arvoista voidaan sanoa kuinka paljon toinen arvo on toista suurempi. Esim: lämpötila celcius-asteina. Kvantitatiivinen.

Suhdeasteikko: jos intervalliasteikon ominaisuuksien lisäksi määriteltynä on yksikäsitteinen mittalukujen absoluuttinen nollapiste, esimerkiksi kunnan veroäyri tai henkilön pituus: nollapiste on 0. Kvantitatiivinen.

Kontrolloidut kokeet ja suorat havainnot

Kontrolloiduissa kokeissa tutkimusaineisto kerätään niin, että tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhteisiin ja selvitetään miten miten kohteet reagoivat muutoksiin.

Suoria havaintoja käyttäessä tutkimusaineisto kerätään niin, että koeolosuhteita ei aktiivisesti muuteta, vaan seurataan miten olosuhteiden muutokset vaikuttavat kohteisiin.

Kummassakin tapauksessa tilastoyksiköihin voi vaikuttaa lisäksi erilaiset **selittävät** ja **sekoittavat tekijät**, joiden vaikutusten kontrollointi on suoria havaintoja tehdessä vaikeampaa!

Puuttuvien selittäjien harhalla tarkoitetaan tilannetta, jossa saatuihin tuloksiin vaikuttaa jokin tekijä, jota ei havaita tai jota ei pystytä mittaamaan ja jonka vaikutusta ei tällöin kyetä kvantifioimaan.

Keskeiset termit: valikoituminen ja satunnaistaminen

Valikoituminen

Valikoitumista tapahtuu, jos otokseen poiminta ei ole riippumatonta tilastoyksikön ominaisuuksista. Tätä kutsutaan valikoitumisharhaksi ja se estää tulosten luotettavan yleistämisen populaatioon.

Satunnaistaminen

Tilastoyksiköiden poimimista populaatiosta otokseen riippumatta muiden yksiköiden poiminnasta tai poimittavien yksiköiden ominaisuuksista. Satunnaistaminen poistaa valikoitumisharhan takaamalla että mahdolliset sekoittavat tekijät ovat jakautuneet tasaisesti tutkittavassa joukossa.

Otantamenetelmät

Otantamenetelmän, joskus myös **otanta-asetelman**, valinta on vahvasti sovellusalakohtainen: käytettävät aineistot ja täten otantamenetelmät määräytyvät pitkälti tehtävän tutkimuksen luonteen perusteella.

Otanta-asetelmalla tarkoitetaan erityisesti otoksen poimintaan käytettyä **satunnaistuksen menetelmää**. Koska tavoitteena on edustava otos, otannan käytäntöön vaikuttaa se, miten todennäköistä kullakin perusjoukon alkiolla on tulla poimituksi otokseen.

Sisältymistodennäköisyys
Sisältymistodennäköisyys kuvaa sitä (tunnettua) todennäköisyyttä, jolla perusjoukon alkio tulee poimituksi otokseen. Merkitään π_k, jolle pätee $0 < \pi_k \leq 1, k = 1, \ldots, N$, kun perusjoukon koko on N alkiota, ts. jokaisella alkiolla on nollaa suurempi sisältymistodennäköisyys.

Yksinkertainen satunnaisotanta (YSO)

Yksinkertaista satunnaisotantaa (YSO) pidetään otannan perusmuotona, jossa jokaisella perusjoukon alkiolla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen, ts. sisältymistodennäköisyys ei riipu tilastoyksikön ominaisuuksista tai otokseen jo valittujen ominaisuuksista. Korjaa siis valikoitumisharhan!

YSO:n toteuttaminen etenee vaiheittain muodostamalla ensin lista kaikista perusjoukon alkioista, joista otanta voidaan suorittaa. Tätä kutsutaan **otantakehikoksi**. Tämän jälkeen alkiot poimitaan otokseen yksi kerrallaan satunnaisesti (arpomalla).

YSO voidaan suorittaa joko **palauttaen** tai **palauttamatta,** jotka poikkeavat siinä miten alkioita kohdellaan sen jälkeen kun ne on poimittu otokseen. Tämä taas vaikuttaa siihen, miten sisältymistodennäköisyydet muuttuvat otannan edetessä! Yksityiskohdat löydät luentomateriaalista.

Systemaattinen otanta

Systemaattisessa, eli tasavälisessä, otannassa perusjoukkoon kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k. alkio. Ei oikeastaan satunnaisotantaa, sillä ei hyödynnä arvontaa!

Potentiaalisen ongelman muodostaa havaintoyksikkölistän mahdollisesti sisältämä säännöllinen jaksollisuus.

Esimerkki: valitaan tehtaan laadunvalvonnassa tuotantolinjalta joka sadas valmistuva tuote laatuarviointiin.

Mikäli alkiot järjestetään satunnaiseen jonoon, on kyseessä kuitenkin vain erilainen tapa toteuttaa yksinkertainen satunnaisotanta!

Ositettu otanta

Joskus tutkimuskohteena oleva perusjoukko koostuu jonkin ominaisuuden suhteen homogeenisista ryhmistä, jotka ovat myös itsessään tutkimuksen kannalta keskeisiä. Tällöin tulee varmistaa, että tutkittava otos on edustava kaikkien olennaisten ryhmien osalta.

Esimerkiksi jos tavoitteena on tutkia jonkin maan erilaisten ja usein hyvin eri kokoisten kieliryhmien taloudellista asemaa ei maan koko populaatioon kohdistettu YSO olisi järkevää, sillä otoskoon pitäisi todennäköisesti olla hyvin suuri että jokaisesta kieliryhmästä saataisiin poimittua edustava otos.

Ositettu otanta ratkaisee ongelman pyrkimällä tunnistamaan tutkimuksen kannalta keskeiset ryhmät ja suorittamalla YSO näiden ryhmien sisältä ja yhdistämällä nämä osaotokset yhdeksi otokseksi.

Ryväsotanta

Paikoin tutkimuskohde voidaan jakaa luonnollisiin ryhmiin eli rypäisiin (eng. *clusters*), jotka indikoivat aineiston luontaista hierarkkista, eli monitasoista- tai asteista rakennetta.

Esimerkiksi koululuokat muodostavat rypäitä koulujen sisällä ja koululaiset ovat alkioita omissa rypäissään.

Ryväsotanta usein motivoidaan tietojen keruun aiheuttamien kustannusten vähentämisellä, sillä rypäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, jolloin jokaista rypästä ei tarvitse erikseen tutkia.

Yksivaiheisessa ryväsotannassa poimitaan joukko rypäitä kaikkien rypäiden joukosta (satunnaisia kouluja), joiden kaikki alkiot tutkitaan (valitun koulun kaikki alkiot). **Kaksivaiheisessa ryväsotannassa** poimitaan vielä satunnaisesti aliryppäät (valitun koulun jotkin luokat) ensimmäisen vaiheen rypäiden joukosta, joista otos poimitaan.

Otannan haasteita kootusti

Tässä tiivistelmässä ohitettiin keskustelu **otoskoon** määrittämisestä. Otoskoko on keskeinen tekijä otoksen edustavuuden kannalta ja siihen palataan vielä myöhemmissä jaksoissa. Yksi otantatutkimuksen uhka on ns. **vastauskato** eli että tutkimuksen kohteita ei tavoiteta tai he kieltäytyvät vastaamasta, jolloin otoskoko pienenee.

Otoskehikolla tarkoitetaan sitä perusjoukon osaa, josta otanta ylipäätään pystytään suorittamaan. **Otoskehikon yli- tai alipeitolla** taas tarkoitetaan tilannetta, jossa otantakehikkoon kuuluu perusjoukkoon kuulumattomia alkioita tai siitä puuttuu osa perusjoukon alkioista.

Poimintaharha: otos ei edusta populaatiota. Vaarana erityisesti silloin, kun otokseen tulleet populaation alkiot ovat valikoituneet tai ovat itse valinneet tiensä otokseen. Aiheutuu myös, kun otoskehikon ali- tai ylipeitto on liian suuri.