SUOMEN TILASTOSEURAN VUOSIKIRJA 2017–2018

ÅRSBOK FÖR STATISTISKA SAMFUNDET I FINLAND 2017–2018

THE YEARBOOK OF THE FINNISH STATISTICAL SOCIETY 2017–2018

2018

Sisältö

Esimiehen palsta

Pauliina Ilmonen

”Niin muuttuu maailma, Eskoni”, kirjoitti Kivi aikoinaan. Suomen tilastoseura on lähes 100-vuotias ja voi kuinka maailma on tämän lähes sadan vuoden aikana muuttunut.

Elämme vaihtoehtoisten totuuksien ja valeuutisten aikaa. Pystymme keräämään ja varastoimaan massiivisia määriä tietoa ja tiedon levittäminen on tänä päivänä globaalistikin hyvin helppoa. Meillä ei kuitenkaan aina ole kykyä, eikä aikaakaan, arvioida tiedon luotettavuutta. Kuka tahansa voi tehdä tilastollista analyysiä ja julkaista löydöksensä. Tämä lisää tutkimuksen läpinäkyvyyttä ja se on hyvä asia. Kuitenkin, jos aineisto ei täytä käytetyn menetelmän vaatimia oletuksia, saattavat analyysin tulokset olla harhaan johtavia tai jopa virheellisiä. Tarvitsemme virallisia luotettavia tilastoja. Tarvitsemme tilastollisten menetelmien osaajia.

Tiedon määrä ja luotettavuus tai sen puute eivät ole ainoita muutoksia. Aineistot, joita käsittelemme ovat hyvin erilaisia kuin aineistot sata vuotta sitten. Tämän päivän aineistot voivat koostua biteistä, funktioista, kuvista, videoista, väreistä tai vaikkapa äänistä. Perinteiset menetelmät, jotka nojaavat esimerkiksi normaalijakauma-oletukseen, eivät sovellu tällaisten aineistojen tarkasteluun. Tarvitsemme uusia menetelmiä, jotka soveltuvat uuden tyyppisten tilastojen analysointiin.

Tekoäly tekee tuloaan. Tekoälyssä ei oikeastaan ole mitään keinotekoista eikä se itsessään ole älykästäkään. Tekoälysovellukset ovat ihmisen luomia. Ne perustuvat (massiivisten) aineistojen analysointiin, algoritmeihin ja matematiikkaan. Toisin sanoen, tekoäly nojaa tilastotieteeseen. Tekoälyä ei tarvitse pelätä. Tekoälysovellukset voidaan nähdä yhtenä tilastotieteen osa-alueena. Perinteiset menetelmät ja tekoälysovellukset elävät symbioosissa. Tekoälysovellukset tarvitsevat tilastoja ja tilastotiedettä. Lisäksi, perinteisiä tilastollisia menetelmiä voidaan käyttää tekoälysovellusten hyvyyden arvioimisessa. Esimerkiksi, jos tekoälysovellus on (vahingossa tai tahallaan) opetettu syrjimään tiettyjä ihmisryhmiä, niin tilastotieteilijä voi olla veljensä tekoälyn vartija.

Oikeastaan elämme tilastojen ja tilastotieteen kulta-aikaa. Meitä tilastotieteilijöitä tarvitaan. Toki nimikkeet muuttuvat. Osa meistä kutsuu itseään nimellä data scientist, osa puhuu datavelhoista, osa meistä kutsuu itseään perinteisesti tilastotieteilijäksi. Tilastoja kerätään enemmän ja enemmän. Niitä opitaan käyttämään päätöksenteossa aina vain paremmin. Me tilastotieteilijät olemme tämän päivän supersankareita.

European meeting of statisticians

European meeting of statisticians (EMS) 2017 järjestettiin 24.—28.7.2018 yhteistyössä Helsingin yliopiston ja Aalto-yliopiston perustieteiden korkeakoulun kanssa. Tapahtumapaikkana oli Helsingin yliopisto. Konferenssi oli menestys. Tapahtumaan osallistui yli 400 tieteentekijää eri maista. Sessioita oli yli 60. Pääpuhujiksi oli houkuteltu professorit Martin Wainwright, Mark Girolami, Alison Etheridge, Alexander Holevo, Gerda Claesken, Hannu Oja, John Aston ja Yann LeCun.

Tilastopäivät

Tilastopäivät järjestettiin Turun yliopistolla 18.—19.5.2017. Tilastopäivien teemana oli Missing Data ja pääpuhujana oli professori Niels Keiding Kööpenhaminan yliopistosta. Tilastopäivillä pohdittiin puuttuvan tiedon ongelmaa sekä teorian näkökulmasta että erilaisten tapausesimerkkien kautta. Konferenssiillallinen järjestettiin Kupittaan paviljongissa. Illallisen aikana keskusteltiin puuttuvan tiedon ongelmista ja pohdittiin tilastotieteen asemaa Suomessa.

Palkinnot

Tilastoseura myönsi Väitöskirjapalkinnon parhaasta vuosina 2013—2016 julkaistusta väitöskirjasta Johan Pensarin työlle Structure Learning of Context-Specific Graphical Models sekä Leo Törnqvist -palkinnot parhaille opinnäytetöille vuosilta 2015—2016 Niko Liétzenin diplomityölle New Approach to Complex Valued ICA: From FOBI to AMUSE ja Santtu Tikan pro gradu -tutkielmalle Kausaalivaikutusten identifiointi algoritmisesti. Palkinnot jaettiin Turussa vietetyillä Tilastopäivillä.

Tilasto-olympialaiset

Suomalaisnuoret toivat voiton kotiin ensimmäisistä tilasto-olympialaisista. Euroopan tilastokilpailu järjestettiin ensimmäisen kerran vuonna 2017. Kilpailu järjestettiin 11 maassa ja osallistujia oli yhteensä yli 11000. Euroopan finaalissa yläkoulusarjan voittoon ylsivät Helsingissä sijaitsevan Pakilan yläkoulun oppilaat Ville Majaniemi, Otto Söderman ja Eero Pohjola. Voittotyö on humoristinen video, joka saa pohtimaan sitä, että tilastot ovat hyödyllisiä niin yhteiskunnalle kuin yksilöillekin. Yläkoulusarjassa hopeaa sai Slovenia ja pronssille tuli Puola. Lukiosarjan voitto tuli Espoossa sijaitsevan Olarin lukion oppilaille Toivo Rannilalle, Asla Heiskaselle ja Olli Hakalalle. Voittotyö on video, joka avaa hauskalla tavalla tilastojen merkitystä poliittisessa ja yhteiskunnallisessa päätöksenteossa. Lukiosarjassa hopeaa sai Kypros ja pronssia Italia. Tilasto-olympialaiset 2018-2019 ovat jo käynnissä! Tilastoseura kannustaa nuoria osallistumaan tilasto-olympialaisiin.

Nuorten seminaari

Tilastoseura ja Biostatistiikan seura järjestivät 20.11.2018 yhteistyössä Aalto-yliopiston perustieteiden korkeakoulun kanssa nuorten seminaarin. Seminaarissa tilastotieteen jatko-opiskelijat tutustuivat toisiinsa ja esittelivät tutkimusaiheitaan.

Iltapäiväseminaarit

Tilastoseura järjesti 30.8.2017 yhteistyössä Terveyden ja hyvinvoinnin laitoksen sekä Jyväskylän yliopiston matematiikan ja tilastotieteen laitoksen kanssa iltapäiväseminaarin Terveyden ja hyvinvoinnin laitoksella. Iltapäiväseminaarin aiheena oli osallistumisen vähyys kyselytutkimuksiin. Seminaarissa pohdittiin muun muassa sitä, että ketkä jättävät vastaamatta terveydentilaan liittyviin kyselytutkimuksiin.

Oikeustieteellinen iltapäiväseminaari järjestettiin 29.1.2018 yhdessä Itä-Suomen yliopiston oikeustieteiden laitoksen kanssa Helsingin yliopiston Kielikeskuksen Juhlasalissa. Iltapäiväseminaarissa pohdittiin muun muassa rangaistusten ankaruuden vaihtelua yksittäisten tuomarien ja syyttäjien välillä.

Vuoden 2018 aikana Tilastoseura järjesti yhdessä Tilastokeskuksen kanssa kaksi iltapäiväseminaaria teemalla ”Mikä on relevanttia ja luotettavaa tietoa ja mistä se tunnistetaan?” Ensimmäinen iltapäiväseminaari pidettiin 12.4.2018. Iltapäiväseminaarissa mietittiin, että miten virallinen tilasto palvelee käyttäjiään ja yhteiskuntaa ja pohdittiin virallisen tilaston vahvuuksia, haasteita ja mahdollisuuksia kehittyä ja vastata käyttäjien tarpeisiin. Toinen seminaari järjestettiin 6.9.2018 Tilastokeskuksen tiloissa. Tämän toisen iltapäiväseminaarin aiheena oli tilastot muuttuvassa, vaihtoehtoisten totuuksien maailmassa.

Odotettavissa

Tilastoseuran järjestämät iltapäiväseminaarit ovat olleet todella suosittuja ja niiden järjestämistä jatketaan. Nuorten seminaari oli innostava ja se järjestetään jatkossa joka vuosi. Kesällä 2019 ei järjestetä tilastopäiviä. Tilastopäivät järjestetään seuraavan kerran vasta kesällä 2020 Suomen Tilastoseuran 100-vuotisjuhlan yhteydessä. Vaikka kesällä 2019 ei järjestetäkään tilastopäiviä, niin Leo Törnqvist -palkinto jaetaan myös kesällä 2019. Seuraava väitöskirjapalkinto jaetaan kesällä 2020 Tilastoseuran 100-vuotisjuhlan yhteydessä.

Sihteerin kommentti

Tommi Mäklin

Maailma tosiaan muuttuu, ja niin muuttuu myös tieteellisen seuran sihteerin toimenkuva. Aloittaessani tilastoseuran sihteerinä 2017 sain erinomaisen perehdytyksen edelliset kaksi vuotta seurasta näyttämön takaa huolehtineelta *Paula Bergmanilta*, mutta enenevissä määrin verkko- ja someaikaan siirtyvän toiminnan vaatimukset ja laajuus pääsivät hieman yllättämään. Oman lusikkansa soppaan työnsivät seuran jäsenyydestä kiinnostuneet kolmatta kotimaista taitavat tilasto-osaajat, joiden palvelemiseksi seuran verkkosivujen ja tiedotuksen uudistamisen urakka laajeni käännöstyöllä englanniksi. Haaveissa siintää vielä seuran ruotsinkielisen jäsenistön parempi huomioiminen.

Kuluneiden kahden vuoden suurimpia muutoksia ovat olleet tilastoseuran lähes kokonaisvaltainen siirtyminen sähköiseen maailmaan niin tiedotuksen kuin hallinnollisten yksityiskohtien saralla sekä näitä seurannut pyrkimys kasvattaa seuran – ja tilastotieteen – yhteiskunnallista näkyvyyttä internetissä ja sosiaalisessa mediassa. Niinikään myös Euroopan unionin vuonna 2018 voimaan tullut tietosuojasetus GDPR aiheutti omat kommervenkkinsä seuran toimintaan. Tilastoseura on myös itse pyrkinyt uudistumaan teknologian omaksumisen saralla, mistä jäsenistölle näkyvimpiä muutoksia ovat sosiaalisen median hyödyntäminen seuran tiedottamisessa ja tämän vuosikirjan ulkoasun yhtenäistäminen sekä päivittäminen taiton siirtyessä toteutettavaksi LaTeX-ladontajärjestelmällä.

Myöhäistietoyhteiskunnan asettamista haasteista on kompastelusta huolimatta selvitty kohtuullisen kunniakkaasti seuran hallituksen osaamisen ja tuen avulla. Erityinen kiitos kuuluu seuraa neljä vuotta luotsanneelle *Jyrki Möttöselle*, joka siirtyi vuoden 2018 alusta hallituksen rivijäseneksi, sekä seuran hallituksen muille lyhyt- ja pitkäaikaisille jäsenille. Tuoreen puheenjohtajan *Pauliina Ilmosen* kaudella on puolestaan saatettu alkuun lukuisia tilastoseuraa kohti tulevaa vieviä hankkeita — näistä lisää seuraavassa vuosikirjassa. Vaikka tulevan ennustaminen on tilastotieteilijöiden lisäksi ollut perinteisesti erinäisten tietäjien ja salatieteiden harjoittajien erikoisalaa, voin tässä silti suurella varmuudella veikata tilastotieteen ja -seuran merkityksen kasvavan yhä suuremmaksi mentäessä kohti seuran satavuotisjuhlia ja toista vuosisataa.

European meeting of statisticians 2017

Pauliina Ilmonen
Chair of the local organizing committee of EMS 2017

In July 2017, Finnish Statistical Society organized, in collaboration with the University of Helsinki and Aalto University School of Science, the 31st European Meeting of Statisticians (EMS). EMS is the main conference in statistics and probability in Europe. It is a biennial conference sponsored by the European regional committee of the Bernoulli Society. The very first EMS conference was organized In Dublin in 1962, and the 31st meeting was held in Helsinki.

EMS 2017 attracted more than 400 participants from different countries and continents. Scientists from different subfield of statistics and probability gathered together and discussed about new ideas and developments. There were eight keynote speakers, 22 invited sessions, 15 topic-contributed sessions, 30 contributed sessions and a poster session. The topics of the presentations varied from the theory of functional data analysis, Bayesian inference, and non-parametric methods to applications in medicine and economics and beyond.

Martin Wainwright gave the opening plenary talk of the conference. His topic was related to pairwise ranking and crowd-sourcing. Mark Girolami gave two forum lectures. He talked about diffusions and dynamics on statistical manifolds for statistical inference. Alison Etheridge gave a special invited lecture about modelling evolution in a spatial continuum. European mathematical society lecture was given by Alexander Holevo. His topic was Quantum Shannon theory. Gerda Claesken gave a special invited lecture about effects of model selection and weight choice on inference, Hannu Oja gave a special invited lecture about scatter matrices and linear dimension reduction, and John Aston gave a special invited lecture about functional object data analysis. The closing lecture was given by Yann LeCun. He talked about deep learning.

EMS 2017 was very successful. Participants were pleased with the program and enjoyed the presentations. On top of that, the weather was beautiful and the participants were excited about the excursions. Sauna, swimming, Finnish forest, open fire and sausages were memorable.

Tilastopäivät 2017 — Missing data

Tilastotieteen keskus, Turun yliopisto
Suomen Tilastoseura
Suomen Biostatistiikan Seura

Turkutime: The Missing Data Song

It's Turkutime,
And Statistics is easy,
Some data are missing,
But randomness is high!

Your data are rich,
Your imputations good-looking,
So hush! Statisticians,
Don't you cry.

Michael Greenacre

Program

Thursday, 18 May

12:00-12:10 Opening words

12:10-13:00 Introduction to missing data — concepts and perspectives
Juha Karvanen (JYU)
Kari Auranen (UTU)

13:00-14:00 Niels Keiding (U Copenhagen)

14:00-14:30 Coffee and tea

14:30-15:30 Short presentations I

1. Henni Pulkkinen (LUKE): Missing information from the Bayesian perspective – examples from fisheries stock assessment
2. Mikko Sillanpää (UO): Mapping genes from a single tail sample of the phenotype distribution by generating pseudo observations
3. Jarno Vanhatalo (HU): Spatio-temporal Gaussian process to detect outbreaks of pests: a case study from the Great Barrier Reef

15:30–16:30 Awards and winners'presentations

16:30–16:45 Coffee and tea

16:45–17:45 Short presentations II

4. Juha Alho (UH): Statistical aspects of Suomi24 discussion forum
5. Reijo Sund (UEF): Are we missing something with complete register data?
6. Mervi Eerola (UTU): Attrition in longitudinal studies

17:45–18:30 Statnet discussion

19:00–23:00 Dinner

Friday, 19 May

9:00-10:00 Niels Keiding (U Copenhagen)

10:00–10:30 Coffee and tea

10:30–12:00 Missing data in health examination surveys
Chair: Jaakko Reinikainen

1. Hanna Tolonen (THL): Selective non-participation in health surveys and available auxiliary information for non-participation adjustment
2. Juho Kopra & Juha Karvanen: Methods for estimating smoking prevalence under selective non-participation
3. Tommi Härkänen (THL): Systematic handling of missing data in complex study designs

12:00–13:00 Lunch

13:00–14:00 How should we evaluate missing information? Comments and critics by the VV club ("Last of the summer wine").
Chair: Elja Arjas

14:15- Statnet discussion

Introduction to missing data – concepts and perspectives

Juha Karvanen University of Jyväskylä
Kari Auranen University of Turku

In this tutorial talk we review some key concepts underlying appropriate treatment of missing data and missing information in statistical inference. We revisit the standard notions of missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) with the aid of technical definitions but also using concrete illustrations. We discuss a recent clarification of missing at random assumptions, based on whether or not the observed missingness pattern is conditioned upon. We mention how the possibility to ignore the missing data mechanism may depend on the assumed statistical paradigm. We review an example where the data are MNAR but the joint distribution can be still estimated in a non-parametric form. Finally, a new vocabulary of missing data terminology in Finnish is introduced.

Web-based enrollment and other types of selection in surveys and studies: consequences for generalizability

Niels Keiding University of Copenhagen

Web-based enrollment in surveys and studies is increasingly attractive as the internet is approaching near-universal coverage and respondents' attitude towards participation in classical modes of study deteriorates. Follow-up is also facilitated by the web-based approach. However, the consequent self-selection raises the question of the importance of representativity when attempting to generalize the results of a study beyond the context in which they were obtained.

In our recent survey of these matters (Keiding and Louis 2016) we described the strong attitude of influential epidemiologists (Miettinen 1985, Rothman et al. 2008) that the methodological focus should be on internal validity in the study sample while scientific generalization should take place within the subject matter areas rather than through formal analysis of external validity of the sample results in larger concrete populations. That view leaves little space for methodological considerations about generalizability beyond the study sample.

The randomized trial is the traditional shining example of internal validity with its efficient confounder control within the study sample. Considerable recent attention has however been devoted to the fact that participation in the sample is voluntary and thus open to self-selection, raising questions about representativity of this sample and the possible consequences for generalizability of the results, particularly under effect heterogeneity. Driven to a large extent by concrete research projects in public health and social sciences, a methodological literature is developing on difficulties about generalizability of effect estimates from randomized studies as well as from observational epidemiological studies.

Our exposition will be divided into three main themes: (1) simple surveys or prevalence studies, assessing the frequency or prevalence of some attitude or

disease condition in a population from its frequency in a sample from this population; (2) generalization of the results from randomized trials to the population in which they were performed and to other populations; and (3) generalization of results from observational studies, again to the population in which they were done, or to other populations. An overarching theme will be synthesizing and "transporting" results. The presentation is based on joint work with Thomas A. Louis.

Missing data or ignored information? Biological processes as a basis of an arrival model for Atlantic salmon smolts

Henni Pulkkinen Natural Resources Institute Finland LUKE

Data available for models of fisheries stock assessment is often sparse and incomplete. In particular, arrival data of migratory species often suffers from breakdowns of monitoring devices or from challenging environmental conditions (e.g. high level of water) that hamper the setup of the monitoring system. A cumulative distribution can be fitted to arrival data predicting abundances for missing dates but such approach easily fails to detect, for example, an early timing of the migration. As an improved method, biologically more realistic arrival model is introduced, using video count data for Utsjoki salmon smolts as a case study. In this model expert knowledge is elicited to connect the timing of the smolt migration and its most important environmental trigger, temperature. Additionally, connection between flow and smolt's migration speed and/or level of water and observation process in the video system can be accounted for. This approach avoids the need to assume any particular functional form for the run curve and can be adjusted to any special characteristics a stock or a monitoring process may have.

Mapping genes from a single tail sample of the phenotype distribution by generating pseudo observations

Mikko Sillanpää University of Oulu

Consider offspring resulting from genetic inbred line cross design in animals or plants. Further, for some reason, Gaussian outcome variable is available among offspring only from a single tail sample of distribution. This is, observations from other tail of Gaussian distribution are missing completely - individuals who have outcome variable missing are also lacking all the covariates. Erroneous analysis occur, if single tail of observations are directly analysed assuming normality. Therefore, we first create pseudo-observations belonging to another tail of the distribution and then analyse the complete sample. This work is from old paper (Sillanpaa & Hoti 2007).

Spatio-temporal Gaussian process to detect outbreaks of pests: a case study from the Great Barrier Reef

Jarno Vanhatalo University of Helsinki

Cyclical outbreaks of pests and diseases can impact the functioning of entire ecosystems. An eminent example is outbreaks of crown-of-thorns starfish (COTS; Acanthaster planci) that cause substantial coral mortality on the Great Barrier Reef (GBR). In order to understand the reasons behind these phenomena we need methods to analyze the spread of pests in space and time. In my talk I will present an analysis of COTS abundance and outbreaks with a Bayesian spatiotemporal Gaussian processes model applied to a long-term survey of the GBR (1985–2014). We assessed the relative increase in COTS abundance beyond that explained by a reef's location and explanatory covariates. Our results confirm that waves of COTS outbreaks originate near Lizard Island (14*67*S) and progress in a northwesterly or southeasterly direction, with the southward wave progressing about 60 km/year. The model reveals several previously unidentified hotspots with high average COTS abundance. The abundance of COTS may also have decreased on reefs protected from fishing after an expansion of protected areas within the GBR Marine Park in 2004, which suggests that closing reefs to fishing may help control COTS. In summary, spatiotemporal Gaussian processes can help identify outbreaks of pests from noisy long-term spatially extensive data which helps managers choose appropriate control strategies. This modelling approach is applicable to other ecosystems where outbreaks of damaging pests and diseases occur.

Statistical aspects of Suomi24 discussion forum

Juha Alho University of Helsinki

Data with social science relevance are rapidly becoming available in digital form. The ease of access to such data will change the way research agendas are formulated. Our first experiences with data from the Suomi24 discussion forum indicate that there can be mechanisms within the data generating process that lead to missing data. Their effect can be larger than the biases caused by self-selection.

Are we missing something with complete register data?

Reijo Sund University of Eastern Finland

Yes, we are always missing something. Data are only an extremely limited representation of reality. Moreover, register data are secondary data - i.e. originally produced for some other purposes than the research in which it will be utilised - meaning that there is no more possibility to tailor data collection so that it would correspond the needs of the research. Available data may or may not be

suitable for the purposes of the research and in virtually any case require a lot of preprocessing before actual analyses. On the other hand, preprocessed register data may be complete in the sense that similar longitudinal event-based data are available for the whole unselected nationwide population so that the problems of drop-out and non-response are no more major issues. In any case, the traditional missing data problem meaning that some variables have missing values for a few fixed time points becomes more or less artificial or irrelevant - reality (and also register-data) are more than a few variables in fixed time points. Why should we talk about missing data at all if we are actually missing most of the reality with any data? Would it be better just to focus on the modelling of reality using available data.

Attrition in longitudinal study of depression

Mervi Eerola University of Turku

A follow-up study of 16-year-old adolescents in Tampere (the TAM project) at ages 22 and 32, initially starting from 2194 individuals (96.7% from the cohort), suffered from attrition in later panels. The non-response pattern and loss of representativeness was studied with inverse probability weighting methods which are common in survey studies, as well as with likelihood-based methods by specifying a longitudinal Bayesian model for the outcome. Both require a longitudinal MAR assumption. Under this assumption, the observed prevalence of depression was well included in the confidence and credibility limits provided by the correction methods. Modelling the missing patterns with pattern-mixture models did not suggest informative missingness. However, modelling the missing mechanism is likely to have poor ability to predict non-response.

Selective non-participation in health surveys and available auxiliary information for non-participation adjustment

Hanna Tolonen National Institute for Health and Welfare THL

Health surveys are used to provide information about health and health determinants of population and population sub-groups for evidence-based policy making, planning and evaluation of prevention programmes and research. Selective non-participation in health surveys may create bias to health indicators. Therefore, it is important to understand profiles of those who don't participate to the health surveys and how their absence may influence derived health indicators.

We know that socio-demographic profile of health survey non-participants tend to include more men, persons from younger age groups or alternatively above 75 years old, un-married, and those having lower education. Non-participants also are shown to have unhealthier lifestyles, i.e. they are more often smokers and heavy alcohol users. There is also evidence that survey non-participants have more health problems, such a mental health, cardiovascular

diseases and cancer. Based on this information, we can say that health survey non-participation is selective on factors of interest.

To evaluate the magnitude of non-participation bias, auxiliary information for non-participants is needed. For health examination surveys conducted by THL (Health 2000/2011, Finrisk study, FinHealth survey, MAAMU survey), different sources of auxiliary information about non-participants has been collected. In some of the studies, short questionnaires or telephone interviews have been conducted on non-participants to obtain at least few data items or re-examination data from previous rounds is available. In all studies, entire sample (participants and non-participants) have been linked to several administrative registers such as causes of death register, hospital discharge information, reimbursement and prescription of medications, cancer register, and socio-demographic information containing e.g. education and profession. This auxiliary information can be used to correct for the effects of non-participation through statistical methods such as weighting, multiple-imputation and Bayesian data augmentation.

NoPaHES web site at http://www.ehes.info/nopahes

Methods for estimating smoking prevalence under selective non-participation

Juho Kopra University of Jyväskylä

Estimation of population statistics from a survey sample under selective non-participation is a challenge. The selective non-participation potentially causes sample statistics, such as averages and prevalences, to be biased. This bias cannot be disposed by adjusting for the background variables but we need external information about the non-participants. We have used two approaches to reduce the bias related smoking prevalence estimates in the FINRISK studies. In the first approach, we use data from a subset of non-participants who answered a questionnaire after a recontact. In the second approach, we utilize survival data on lung cancer and COPD as an indirect way to gain information about smoking prevalence. Multiple imputation is used in the first approach and Full Bayesian analysis in the second approach.

Systematic handling of missing data in complex study designs

Tommi Härkänen University of Turku
Juha Karvanen University of Jyväskylä
Hanna Tolonen National Institute for Health and Welfare THL
Risto Lehtonen University of Helsinki
Kari Djerf Statistics Finland
Teppo Juntunen National Institute for Health and Welfare THL
Seppo Koskinen National Institute for Health and Welfare THL

Correction for the effects of non-participation in surveys requires measures both prior to the data collection and after it. We compare statistical methods to reduce the effects of non-participation. We also assess if repeated measures survey data using the Health 2000 and 2011 surveys (BRIF8901) can provide more accurate results than the cross-sectional Health 2011 survey data alone. All sample members were micro-merged with national registries. Increased non-participation from 2000 to 2011 was a major issue. We apply a systematic approach to the practical and comprehensive handling of missing data motivated by our experiences of analyzing longitudinal survey data. Model assumptions involved in the complex sampling design, repeated measurements design, non-participation mechanisms and associations are presented graphically using methodology previously defined as a causal model with design i.e. a functional causal model extended with the study design.

We have compared inverse probability weighting (IPW), multiple imputation (MI) and doubly robust (DR) to handle missing data in three register-based health outcomes. We found that MI removed almost all differences between full sample and estimated prevalences. The IPW removed more than half and the DR method about 60% of the differences. Inclusion of the baseline information collected in 2000 improved the accuracy of the estimates and reduced their standard errors compared to results based on the cross-sectional data alone. These findings encourage us to conduct repeated measures populations surveys also to estimate cross-sectional population statistics since decreasing participation rates are a major problem in population surveys worldwide.

Väitöskirjapalkinto

Structure learning of context-specific graphical models

Johan Pensar
Åbo Akademi University
Department of Mathematics and Statistics

Abstract

The underlying problem considered in the thesis is modeling the joint distribution of a collection of categorical variables. For this purpose, a new family of context-specific graphical models is introduced, and the main emphasis is on learning the structure of the models from data. This summary provides a high-level overview of the material covered by the thesis. For more details about the models and learning methods, the reader is referred to the cited papers. The thesis is available online at http://urn.fi/URN:ISBN:978-952-12-3413-2.

Introduction

Probabilistic models provide a general tool for modeling real-world systems where there is a significant amount of uncertainty involved. Here, we focus on the class of (probabilistic) graphical models, which are used for modeling complex joint distributions over a set of discrete variables. A compact representation of a potentially very high-dimensional distribution is achieved by exploiting structure in the distribution, corresponding to statements of conditional independence. One of the core ideas behind graphical models is to use a graph structure to compactly encode the dependence structure over the variables.

During the last few decades, graphical models have received considerable attention by the statistics and computer science communities (Koller and Friedman, 2009). Despite their wide adoption, the restrictions implied by conditional independence have been recognised to be unnecessarily coarse in certain situations, resulting in the development of more flexible models (Geiger and Heckerman, 1996; Højsgaard, 2003; Boutilier et al., 1996; Friedman and Goldszmidt, 1996; Chickering et al., 1997; Poole and Zhang, 2003; Corander, 2003). In particular, Boutilier et al. (1996) formalised the notion of context-specific independence (CSI), which is a natural generalisation of conditional independence.

By including CSI into the graphical model framework, it is possible to obtain more accurate model structures which still enjoy a sound independence-based interpretation.

The aim of the thesis is to develop general classes of CSI-based graphical models and, since the mere existence of complex models is of limited practical use, the main focus is on developing methods for learning the structure of the models from data. This summary provides a high-level overview of the material covered by the four articles included in the thesis (Pensar et al., 2015; Pensar et al., 2016; Pensar et al., 2017a; Pensar et al., 2017b). In Section 5.1.2, we begin by briefly describing the concept of graphical models. In Section 5.1.3, we introduce the notion of CSI and explain its role in context-specific graphical models. In Section 5.1.4, we briefly explain how Bayesian score functions, originally developed for learning the structure of traditional models, can be modified to be used on context-specific models. Finally, in Section 5.1.5, we provide some concluding remarks.

Graphical models

We consider a set of d random categorical variables $X = \{X_1, \ldots, X_d\}$. Each variable X_j takes on values from a finite set of outcomes represented by $\mathcal{X}_j = \{0, 1, \ldots, r_j - 1\}$, where r_j is the number of outcomes. We let $V = \{1, \ldots, d\}$ denote the indices of the variables. For a subset $S \subseteq V$, we denote the corresponding variables by X_S. We use $p(X)$ to denote the distribution over X, whereas $p(x)$ is shorthand for the probability $p(X = x)$.

The purpose of graphical models is to represent a joint distribution over X in an efficient and compact manner. Even in the case of binary variables, a naive representation would require as many as $2^d - 1$ free parameters to fully specify a joint distribution over d variables. It is easy to realise that such a representation quickly becomes impractical as the number of variables is increased. To overcome this problem, graphical models break down the joint distribution into smaller more manageable parts by exploiting statements of *conditional independence*.

Definition 1. *Conditional Independence*
Let A, B, S be three disjoint subsets of V. We say that X_A is conditionally independent of X_B given X_S if

$$p(x_A \mid x_B, x_S) = p(x_A \mid x_S)$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_S) > 0$. This is denoted by $X_A \perp X_B \mid X_S$. If $S = \emptyset$, then $X_A \perp X_B$ is reduced to marginal independence between the two sets of variables.

In general, it is not practical to represent the dependence structure of the variables in the form of a list of independence statements. Instead, the dependence structure of a graphical model is encoded by a graph structure. The graph $G = (V, E)$ consists of *nodes* (or *vertices*), $V = \{1, \ldots, d\}$, representing the variables in the model, and *edges*, $E \subset V \times V$, representing direct dependences between the variables. Similarly, lack of edges implies statements of conditional independence.

The graph offers an intuitive and compact way of illustrating the dependence structure to a human user. Moreover, it enables the use of graph theory when

(a) (b)

Figure 5.1.1: Example graphs of (a) a Bayesian network and (b) a Markov network over four variables.

designing inference algorithms for the model. There are two main families of graphical models, which are characterised by their type of graph; *Bayesian networks* (directed graphical models) and *Markov networks* (undirected graphical models).

Bayesian networks

The basis of the Bayesian network formulation is a *directed acyclic graph (DAG)*. We denote a directed edge by $(i \to j)$, where $i, j \in V$. The edge set E must satisfy the *acyclicity* property; when starting from a node j, it is not possible to return to j following the direction of the edges. The *parents* of node j are denoted by $pa(j)$ and defined as all nodes from which there is a directed edge to j, that is, $pa(j) = \{i \in V : (i \to j) \in E\}$. Finally, the descendants of node j are all nodes that can be reached from j following the direction of the edges.

The dependence structure encoded by a DAG can be characterised by the directed local Markov property, which states that each variable is conditionally independent of its non-descendants given its parents. In addition, these local independences imply a collection of non-local independences. All independences encoded by a Bayesian network can be verified directly from the graph using a graph-theoretic criterion known as *d-separation* (Koller and Friedman, 2009).

As justified by the directed local Markov property, the joint distribution of a Bayesian network factorises over the DAG into node-wise *conditional probability distributions (CPDs)*:

$$p(X_1, \ldots, X_d) = \prod_{j \in V} p(X_j \mid X_{pa(j)}).$$

For example, the factorisation according to the DAG in Figure 5.1.1a is

$$p(X_1, X_2, X_3, X_4) = p(X_1)p(X_2)p(X_3 \mid X_1, X_2)p(X_4 \mid X_3).$$

In general, for a variable X_j, a separate CPD is specified for each distinct parent configuration $x_{pa(j)} \in \mathcal{X}_{pa(j)}$.

Markov networks

The dependence structure of a Markov network is represented by an undirected graph. We denote an undirected edge by $(i - j)$, where $i, j \in V$. A *clique* in a graph is defined as a subset of nodes for which all pairs of nodes are connected by an edge. A clique is defined as *maximal* if no additional node can be added to the clique without violating the clique criterion. The *Markov blanket* of a

node j is denoted by $mb(j)$ and defined as all nodes which are connected to j by an edge, that is, $mb(j) = \{i \in V : (i - j) \in E\}$.

The independence statements encoded by a Markov network can be read off the graph using the standard graph separation criterion. Equivalently, assuming a positive joint distribution, the dependence structure can be characterised by the local Markov property, which states that each variable is conditionally independent of the rest of the network given its Markov blanket.

The (positive) joint distribution of a Markov network factorises over the maximal cliques in the undirected graph. A common parameterisation of the distribution is in form of a log-linear model (Whittaker, 1990):

$$\log p(x_1, \ldots, x_d) = \sum_{A \subseteq V} \phi_A(x),$$

where the ϕ-terms are real-valued coordinate projection functions, $\phi_A(x) = \phi_A(x_A)$, which satisfy the graph-related constraint:

$$\phi_A(\cdot) = 0 \text{ if } \{i, j\} \subseteq A \text{ for some } (i - j) \notin E.$$

Furthermore, to avoid an over-parameterisation, we also assume that

$$\phi_A(x_A) = 0 \text{ if } x_j = 0 \text{ for any } j \in A.$$

As an example, the log-linear parameterisation associated with the graph in Figure 5.1.1b is

$$\begin{aligned} \log p(x_1, x_2, x_3, x_4) = \phi_\varnothing + \phi_1(x) + \phi_2(x) + \phi_3(x) + \phi_4(x) \\ + \phi_{1,2}(x) + \phi_{1,3}(x) + \phi_{2,3}(x) + \phi_{3,4}(x) + \phi_{1,2,3}(x), \end{aligned} \tag{5.1}$$

where the $\phi_\varnothing$-term serves as a normalising constant for the distribution. In contrast to a Bayesian network, the factorisation of the joint distribution is not “genuine" for (non-chordal) Markov networks, since the model parameters are connected through the normalising constant, also known as the partition function.

Context-specific graphical models

It has been noticed that conditional independence alone can be unnecessarily stringent for modeling real-world phenomena. In an attempt to loosen the restrictions of traditional graphical models, the notion of *context-specific independence (CSI)* has emerged (Boutilier et al., 1996; Friedman and Goldszmidt, 1996; Geiger and Heckerman, 1996; Poole and Zhang, 2003; Corander, 2003; Højsgaard, 2003; Nyman et al., 2014; Nyman et al., 2015). CSI is a natural generalisation of conditional independence and it was formalised by Boutilier et al. (1996) for the purpose of capturing regularities in the CPDs of Bayesian networks by using context trees.

Definition 2. *Context-Specific Independence*
Let A, B, C, S be four disjoint subsets of V. We say that X_A is contextually independent of X_B given X_S and the context $X_C = x_C$ if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S)$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_C, x_S) > 0$. This will be denoted by $X_A \perp X_B \mid x_C, X_S$.

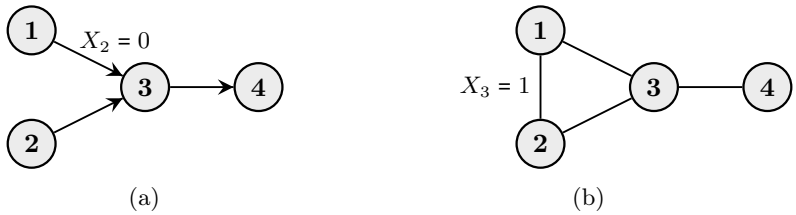


Figure 5.1.2: Example labeled graphs of (a) a context-specific Bayesian network and (b) a context-specific Markov network over four binary variables.

When comparing Definitions 1 and 2, it is easy to realise that CSI is a more specific form of conditional independence that only holds in part of the outcome space of the conditioning variables:

$$X_A \perp X_B \mid x_C, X_S \text{ for all } x_C \in \mathcal{X}_C \quad \Leftrightarrow \quad X_A \perp X_B \mid X_C, X_S.$$

As explained in the next two sections, certain types of local CSIs can naturally be incorporated into the graphical model framework. To encode such additional restrictions in a single graphical structure, we attach labels to the edges of the graph, as originally proposed by Corander (2003). Essentially, a label will encode a context for which the influence of the associated edge "vanishes". Although the theory holds for general categorical variables, for simplicity, we will focus on binary variables in the given examples.

Context-specific independence in Bayesian networks

The conditional independence assumptions made by a Bayesian network enables a compact factorisation of the joint distribution. To further refine the factorisation, the concept of local CSI statements has been proposed and investigated by numerous authors (Boutilier et al., 1996; Friedman and Goldszmidt, 1996; Poole and Zhang, 2003). By local, we refer to a statement concerning the relation between a node and its parents:

$$X_j \perp X_{pa(j)\setminus C} \mid x_C \text{ where } C \subset pa(j) \text{ (Definition 3, Pensar et al. 2015).} \quad (5.2)$$

Local CSI statements are particularly well-suited for the Bayesian network parameterisation, since they imply that certain CPDs are identical. More specifically, the statement in (5.2) implies that

$$p(X_j \mid x_{pa(j)\setminus C}, x_C) = p(X_j \mid x'_{pa(j)\setminus C}, x_C)$$

for all $x_{pa(j)\setminus C}, x'_{pa(j)\setminus C} \in \mathcal{X}_{pa(j)\setminus C}$. Since identical CPDs need only be specified once, the number of necessary model parameters can be reduced accordingly. This can be viewed as partitioning the outcome space of the parents into classes of configurations such that the conditional distribution of the node is invariant for parent configurations belonging to the same class.

As an example, consider the labeled DAG in Figure 5.1.2a. The label on edge $(1 \to 3)$ encodes a local CSI of the form:

$$X_1 \perp X_3 \mid X_2 = 0.$$

In terms of restrictions on the distribution, this implies that $(X_1, X_2) = (0, 0)$ and $(X_1, X_2) = (1, 0)$ induce identical conditional distributions for $X_3 \mid X_{pa(3)}$.

Consequently, the CSI will imply a partitioning of the outcome space of the parental variables $X_{pa(3)} = (X_1, X_2)$:

$$\{(0,0),(0,1),(1,0),(1,1)\} \rightarrow \{(0,1), \begin{Bmatrix}(0,0)\\(1,0)\end{Bmatrix}, (1,1)\}, \tag{5.3}$$

reducing the number of distinct CPDs by one, which in the binary case also reduces the number of free model parameters by one.

Context-specific independence in Markov networks

The notion of CSI has also been investigated as a means for improving the flexibility of Markov networks (Højsgaard, 2003; Corander, 2003; Nyman et al., 2014; Nyman et al., 2015). In particular, Nyman et al. (2015) developed the classes of *(decomposable) stratified graphical models (SGM)*. To facilitate model learning, certain restrictions were imposed on the SGMs. In Pensar et al. (2017a), we introduce a class of general context-specific Markov networks, which subsumes the SGM class.

The context (or label) of an each edge in a context-specific Markov network is specified by the *common neighbours* of the edge nodes. The common neighbours with respect to an undirected edge $(i - j)$ are denoted and defined by $cn(i,j) = mb(i) \cap mb(j)$, and the corresponding local CSI statements are of the form

$$X_i \perp X_j \mid x_{cn(i,j)}, X_{V \setminus \{cn(i,j) \cup \{i,j\}\}} \text{ (Definition 2, Pensar et al. 2017a).} \tag{5.4}$$

Using the common neighbours to specify an edge context is proven to be a natural condition in terms of the generality of the models (Section 2.2, Pensar et al. 2017a). Following a similar reasoning as previous works (Corander, 2003; Nyman et al., 2015), we show that CSI statements of type (5.4) imply certain linear restrictions on the log-linear parameters (Proposition 2, Pensar et al. 2017a).

As an example, consider the labeled undirected graph in Figure 5.1.2b. As defined in (5.4), the label on edge $(1 - 2)$ encodes a local CSI of the form

$$X_1 \perp X_2 \mid X_3 = 1, X_4.$$

Moreover, according to Proposition 2 in Pensar et al. (2017a), the above CSI will imply the following restriction(s) on the log-linear parameters in (5.1):

$$\phi_{1,2}(x) + \phi_{1,2,3}(x) = 0,$$

reducing the number of free model parameters by one (in the binary case).

Structure learning

In this section, we consider the problem of learning the structure of a model from data, commonly known as structure learning. Structure learning is a challenging problem already for traditional graphical models, and bringing in CSI further complicates the matter. Still, the potential gain of a more flexible model structure is a better fit to the data without inducing redundant model parameters.

Here, we focus on score-based structure learning, which can be framed as an optimisation problem. Firstly, this requires a score function by which the plausibility of each possible structure can be evaluated. Secondly, to find high-scoring

networks, this also requires a search algorithm, since an exhaustive evaluation is in general infeasible due to the vast model space. In this summary, we will focus on the former. More specifically, we will briefly describe two Bayesian score functions, which can be used to infer the structure of the context-specific models in Section 5.1.3.

A Bayesian score for structure learning

The Bayesian framework provides a very natural approach for assessing the plausibility of a network structure given a set of observed data. By a data set $\mathbf{x}$, we refer to a complete data set consisting of n i.i.d. joint observations, assumed to have been generated from a model within the considered model class. In the Bayesian approach, the plausibility of a graph G is assessed by the unnormalised conditional probability of the graph given the data $\mathbf{x}$:

$$p(G \mid \mathbf{x}) \propto p(\mathbf{x} \mid G)p(G),$$

where $p(\mathbf{x} \mid G)$ is the marginal likelihood under the given graph and $p(G)$ is the prior probability of the graph. The marginal likelihood is evaluated by

$$p(\mathbf{x} \mid G) = \int p(\mathbf{x} \mid G, \theta) f(\theta \mid G) d\theta, \tag{5.5}$$

where $p(\mathbf{x} \mid G, \theta)$ is the likelihood function under the given graph and $f(\theta \mid G)$ is a prior over the model parameters. The marginal likelihood evaluates how well the observed data fits the given model structure, while implicitly controlling for the complexity of the model. The structure prior $p(G)$ has been given less attention in the structure learning literature, for example, it is common to assume a uniform prior. However, experiments have shown that a sparsity-inducing prior is typically necessary to deal with the added flexibility of context-specific models (Pensar et al., 2015; Pensar et al., 2016; Pensar et al., 2017a).

In the following sections, we will describe how the marginal likelihood and an approximate version of it can be evaluated in closed form for context-specific Bayesian networks and Markov networks, respectively.

Marginal likelihood for context-specific Bayesian networks

Under certain assumptions, listed by Heckerman et al. (1995), the integral in (5.5) can be solved analytically for Bayesian networks. This results in a closed-form expression of the form:

$$p(\mathbf{x} \mid G) = \prod_{j \in V} p(\mathbf{x}_j \mid \mathbf{x}_{pa(j)}) = \prod_{j \in V} \prod_{u \in \mathcal{X}_{pa(j)}} p(\mathbf{x}_j \mid X_{pa(j)} = u), \tag{5.6}$$

where $p(\mathbf{x}_j \mid X_{pa(j)} = u)$ is the standard marginal likelihood under a categorical distribution, $p(X_j \mid X_{pa(j)} = u)$, and a Dirichlet parameter prior (Buntine, 1991; Cooper and Herskovitz, 1992; Heckerman et al., 1995).

Conveniently, the marginal likelihood can readily be modified to also cover context-specific Bayesian networks (Friedman and Goldszmidt, 1996; Chickering et al., 1997; Pensar et al., 2015; Pensar et al., 2016). The key thing to realise is that local CSIs imply a partitioning of the parent outcome space, as explained in Section 5.1.3. In other words, the parent configurations, which are denoted by u in (5.6), are replaced by the classes forming the parent partition.

As an example, consider the labeled DAG in Figure 5.1.2a. At the node level, the marginal likelihood factorises similarly as for the DAG in Figure 5.1.1a:

$$p(\mathbf{x} \mid G) = p(\mathbf{x}_1)p(\mathbf{x}_2)p(\mathbf{x}_3 \mid \mathbf{x}_{1,2})p(\mathbf{x}_4 \mid \mathbf{x}_3).$$

Moreover, the local scores of nodes 1, 2 and 4 are evaluated according to (5.6), since the nodes do not contain any labeled incoming edges. However, to account for the CSI encoded by the labeled edge $(1 \to 3)$, the local score of node 3 is now calculated according to the partition in (5.3), where $(X_1, X_2) = \{(0,0),(1,0)\}$ should be interpreted as the context in which

$$(X_1, X_2) = (0,0) \vee (X_1, X_2) = (1,0).$$

For more details, see Pensar et al. (2015).

Marginal pseudo-likelihood for context-specific Markov networks

Due to the partition function, likelihood-based scores are in general intractable for non-chordal Markov networks. For this reason, alternative objective functions have been proposed, with the perhaps most popular being the *pseudo-likelihood* (Besag, 1975). The pseudo-likelihood approximates the likelihood by a product of conditional likelihoods over each individual node given the remaining variables, or under a fixed graph, given the Markov blanket of the node.

In Pensar et al. (2017b), we introduce the *marginal pseudo-likelihood (MPL)* as an alternative Bayesian-type score for Markov networks:

$$\hat{p}(\mathbf{x} \mid G) = \prod_{j \in V} p(\mathbf{x}_j \mid \mathbf{x}_{mb(j)}) = \prod_{j \in V} \prod_{u \in \mathcal{X}_{mb(j)}} p(\mathbf{x}_j \mid X_{mb(j)} = u), \tag{5.7}$$

which is similar to the marginal likelihood for Bayesian networks (5.6); however, the parent configurations are replaced by Markov blanket configurations. Similar to the marginal likelihood, the MPL is consistent in the large sample limit, that is, the correct network structure will obtain the highest score as $n \to \infty$ (Theorem 1, Pensar et al. 2017b).

As an example of the factorisation of the MPL score, the undirected graph in Figure 5.1.1b is evaluated by

$$\hat{p}(\mathbf{x} \mid G) = p(\mathbf{x}_1 \mid \mathbf{x}_{2,3})p(\mathbf{x}_2 \mid \mathbf{x}_{1,3})p(\mathbf{x}_3 \mid \mathbf{x}_{1,2,4})p(\mathbf{x}_4 \mid \mathbf{x}_3).$$

From a computational perspective, the factorisation property makes the MPL an attractive objective function for search algorithms based on single edge changes (Section 4, Pensar et al. 2017b).

In Pensar et al. (2017a), we extended the scope of the MPL to also cover context-specific Markov networks. The key innovation lies in the observation that the CSI statements of a context-specific Markov network can be accounted for by the MPL by partitioning the outcome space of the Markov blankets in a similar way as the outcome space of the parents in a context-specific Bayesian network (for more details, see Section 3.2, Pensar et al. 2017a). The modified MPL score was shown to be consistent for learning the model structure of context-specific Markov networks (Theorem 1, Pensar et al. 2017a)

Concluding remarks

The notion of context-specific independence (CSI) has been proposed as a means to generalise probabilistic graphical models by allowing for more flexible dependence structures. We have further pursued this idea by building up the theory of the class of context-specific graphical models, in which CSI is included as part of the network structure. The main emphasis of the thesis is on learning the structure of such models from data. To enable efficient structure learning, the scope of Bayesian score functions were extended to also cover context-specific models. Numerical experiments on synthetic and real-world data have shown that the increased flexibility of context-specific structures can more accurately capture the dependence structure among a set variables and thereby improve the predictive accuracy of the models (Pensar et al., 2015; Pensar et al., 2016; Pensar et al., 2017a).

Acknowledgements

I would like to thank all co-authors of the articles included in the thesis, in particular, a special thanks goes to Henrik Nyman and my PhD supervisor Jukka Corander.

Bibliography

[1] J. Whittaker. *Graphical models in applied multivariate statistics*. Chichester: Wiley, 1990.

[2] J. Besag. "Statistical analysis of non-lattice data". In: *Journal of the Royal Statistical Society, Series D (The Statistician)* 24 (1975), pp. 179–195.

[3] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. "Context-specific independence in Bayesian networks". In: *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*. 1996, pp. 115–123.

[4] W. Buntine. "Theory refinement on Bayesian networks". In: *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1991, pp. 52–60.

[5] D. M. Chickering, D. Heckerman, and C. Meek. "A Bayesian approach to learning Bayesian networks with local structure". In: *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*. 1997, pp. 80–89.

[6] G.F. Cooper and E. Herskovitz. "A Bayesian Method for the induction of probabilistic networks from data". In: *Machine Learning* 9 (1992), pp. 309–347.

[7] J. Corander. "Labelled graphical models". In: *Scandinavian Journal of Statistics* 30 (2003), pp. 493–508.

[8] N. Friedman and M. Goldszmidt. "Learning Bayesian networks with local structure". In: *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*. 1996, pp. 252–262.

[9] D. Geiger and D. Heckerman. "Knowledge representation and inference in similarity networks and Bayesian multinets". In: *Artificial Intelligence* 82 (1996), pp. 45–74.

[10] D. Heckerman, D. Geiger, and D.M. Chickering. "Learning Bayesian networks: The combination of knowledge and statistical data". In: *Machine Learning* 20 (1995), pp. 197–243.

[11] S. Højsgaard. "Split models for contingency tables". In: *Computational Statistics & Data Analysis* 42 (2003), pp. 621–645.

[12] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[13] H. Nyman, J. Pensar, T. Koski, and J. Corander. "Context-specific independence in graphical log-linear models". In: *Computational Statistics* (2015). DOI: 10.1007/s00180-015-0606-6.

[14] H. Nyman, J. Pensar, T. Koski, and J. Corander. "Stratified graphical models - context-specific independence in graphical models". In: *Bayesian Analysis* 9.4 (2014), pp. 883–908.

[15] J. Pensar, H. Nyman, and J. Corander. "Structure Learning of Contextual Markov networks using marginal pseudo-likelihood". In: *Scandinavian Journal of Statistics* 44 (2017), pp. 455–479.

[16] J. Pensar, H. Nyman, T. Koski, and J. Corander. "Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models". In: *Data Mining and Knowledge Discovery* 29.2 (2015), pp. 503–533.

[17] J. Pensar, H. Nyman, J. Lintusaari, and J. Corander. "The role of local partial independence in learning of Bayesian networks". In: *International Journal of Approximate Reasoning* 69 (2016), pp. 91–105.

[18] J. Pensar, H. Nyman, J. Niiranen, and J. Corander. "Marginal Pseudo-Likelihood Learning of Discrete Markov Network Structures". In: *Bayesian Analysis* 12.4 (2017), pp. 1195–1215.

[19] D. Poole and N.L. Zhang. "Exploiting contextual independence in probabilistic inference". In: *Journal of Artificial Intelligence Research* 18 (2003), pp. 263–313.

Leo Törnqvist -palkinto

Kausaalivaikutusten identifiointi algoritmisesti

Santtu Tikka
Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos

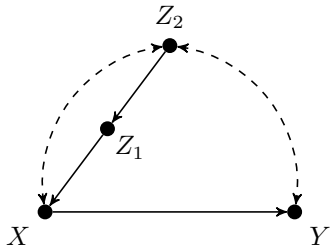
Tiivistelmä

Kausaalisuutta on tilastotieteessä perinteisesti lähestytty kokeellisten tutkimusten avulla. Monissa tilanteissa kokeellisen tutkimuksen toteuttaminen ei kuitenkaan ole mahdollista esimerkiksi käytännöllisistä tai eettisistä syistä johtuen. Tällöin on turvauduttava havainnoiviin tutkimuksiin. Tutkielmassa tarkastellaan Judea Pearlin kehittämää kausaalimallia ja kausaalilaskentaa, joiden avulla voidaan vastata moniin kausaalisuutta koskeviin kysymyksiin myös havainnoivista tutkimuksista saatujen aineistojen avulla. Tutkielma on saatavilla verkossa (http://urn.fi/URN:NBN:fi:jyu-20150330152) ja sen pohjalta on myös julkaistu artikkeli (https://www.jstatsoft.org/article/view/v076i12).

Kausaalimallit

Kausaalisuuden tarkasteluun on tilastotieteen historiassa esitetty lukuisia toisiaan muistuttavia lähestymistapoja, kuten Neyman–Rubin -kausaalimalli (Rubin, 1974) ja rakenneyhtälömallit (Kline, 1998). Nämä menetelmät muistuttavat formalismiltaan regressioanalyysia, eikä mallien kausaalinen tulkinta ole aina suoraviivaista. Graafisten mallien ja Bayes-verkkojen yleistyessä osoittautui, että graafiteoria tarjosi sopivan lähtökohdan myös kausaalisten kysymysten käsittelyyn. Tutkielmassa käsitellään kausaalimallia, jonka kehitti Pearl (1995) sekä malliin olennaisesti liittyvän kausaalivaikutusten identifioituvuusongelman algoritmista ratkaisua, jonka kehittivät Shpitser ja Pearl (2006b).

Kausaalimallin tarkoituksena on esittää kiinnostuksen kohteena olevien muuttujien väliset yhteydet selkeästi ja täsmällisesti. Myös kausaalisuuden suunta on tärkeä ottaa huomioon, sillä syy ei voi edeltää seurausta. *Graafit* ja erityisesti *suunnatut silmukattomat graafit* ovat tärkeitä työvälineitä näiden käsitteiden formalisoimiseksi. Suunnattu silmukaton graafi on pari $G = \langle \mathbf{V}, \mathbf{E} \rangle$, missä



Kuva 6.1.1: semi-Markov-graafi. Havaitsemattomat muuttujat on kuvattu kaksisuuntaisilla särmillä

joukko $\mathbf{V}$ koostuu graafin G *solmuista* ja joukko $\mathbf{E}$ graafin G *särmistä*. Särmät ovat järjestettyjä pareja (V_i, V_j), joille $V_i, V_j \in \mathbf{V}$. *Polku* on järjestetty jono särmiä, jossa seuraavan särmän alkusolmu on edellisen särmän loppusolmu, eli:

$$(V_1, V_2), (V_2, V_3), \ldots, (V_{n-1}, V_n).$$

Suunnatussa silmukattomassa graafissa ei ole olemassa sellaista polkua, joka alkaa tietystä solmusta ja päättyy samaan solmuun, toisin sanoen $V_1 \neq V_n$ on voimassa kaikille graafin poluille.

Graafit eivät yksinään riitä kausaalisuuden malliksi, sillä niiden avulla ei voida ottaa huomioon tilastollista epävarmuutta eikä antaa tulkintaa särmien esittämille kausaalisille yhteyksille. Tätä varten määritellään kausaalimalli (Pearl, 2009). *Probabilistinen kausaalimalli* on nelikko

$$\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle,$$

missä:

1. $\mathbf{U}$ on joukko havaitsemattomia taustamuuttujia, jotka määräytyvät mallin ulkopuolisista tekijöistä.
2. $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$ on joukko havaittuja muuttujia, jotka määräytyvät muiden havaittujen ja havaitsemattomien muuttujien perusteella.
3. $\mathbf{F} = \{f_{V_1}, f_{V_2}, \ldots, f_{V_n}\}$ on sellainen joukko funktioita, että jokainen f_{V_i} on kuvaus joukolta $\mathbf{U} \cup (\mathbf{V} \setminus \{V_i\})$ joukolle V_i.
4. $P(\mathbf{U})$ on muuttujien $\mathbf{U}$ yhteisjakauma.

Kausaalimallia vastaava graafi G sisältää solmun jokaista havaittua ja havaitsematonta muuttujaa kohden. Solmusta V_i kulkee särmä solmuun V_j mikäli solmu V_i on osa funktion f_{V_j} määrittelyjoukkoa. Tutkielmassa keskitytään *semi-Markov-kausaalimalleihin*, joita vastaavissa graafeissa jokaisesta havaitsemattomasta muuttujasta lähtee korkeintaan kaksi särmää havaittuihin muuttujiin ja havaitsemattomien muuttujien välillä ei ole särmiä. Tällöin havaitsemattomat muuttujat kuvataan yleensä kaksisuuntaisina särminä. Kaikki kausaalimallit voidaan palauttaa semi-Markov-kausaalimalleiksi *latentin projektion* avulla (Verma, 1993; Tian ja Pearl, 2003). Kuvassa 6.1.1 on esimerkki semi-Markov-kausaalimallia vastaavasta graafista.

Sopivien oletusten ollessa voimassa, voidaan havaittujen muuttujien välisiä ehdollisia riippumattomuuksia tarkastella suoraan kausaalimallia vastaavan graafin avulla. Graafin G ja todennäköisyysjakauman P yhteyttä kuvaa

d-separaatio (Pearl, 1995; Pearl, 2009; Koller ja Friedman, 2009). Polku H on solmujoukon $\mathbf{Z}$ d-separoima graafissa G, jos ja vain jos

1. H sisältää ketjun $I \to M \to J$ tai haarukan $I \leftarrow M \to J$, missä $M \in \mathbf{Z}$ ja $I, J \in \mathbf{V}$.

2. H sisältää käänteisen haarukan $I \to M \leftarrow J$, missä yksikään solmun M jälkeläisistä ei kuulu joukkoon $\mathbf{Z}$ graafissa G solmu M mukaan lukien, ja $I, J \in \mathbf{V}$.

Jos erilliset muuttujajoukot $\mathbf{X}$ ja $\mathbf{Y}$ ovat solmujoukon $\mathbf{Z}$ d-separoimia graafissa G, niin $\mathbf{X}$ ja $\mathbf{Y}$ ovat ehdollisesti riippumattomia ehdolla $\mathbf{Z}$ graafissa G (jokaisen yhteensopivan jakauman P suhteen).

Kausaalivaikutukset ja identifioituvuus

Kausaalimallin avulla on mahdollista tarkastella kuinka sen kuvaamat muuttujien väliset funktionaaliset suhteet muuttuvat ulkopuolisten toimenpiteiden eli *interventioiden* seurauksena. Formaalisti interventiot määritellään $\mathrm{do}(\cdot)$-operaattorin avulla. Operaattori asettaa argumenttiinsa liittyvät funktiot vakiofunktioiksi. Esimerkiksi interventio $\mathrm{do}(\mathbf{X} = \mathbf{x})$ asettaa kaikkien joukon $\mathbf{X}$ muuttujia vastaavat funktiot sellaisiksi, että ne tuottavat aina vakioarvon $\mathbf{x}$. Interventio kausaalimalliin M tuottaa siten uuden *alimallin* $M_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle$. Muuttujajoukon $\mathbf{Y} \subset \mathbf{V}$ jakaumaa syntyneessä alimallissa kutsutaan intervention $\mathrm{do}(\mathbf{X} = \mathbf{x})$ *kausaalivaikutukseksi* muuttujajoukkoon $\mathbf{Y}$, josta käytetään merkintää $P(\mathbf{Y} \mid \mathrm{do}(\mathbf{X} = \mathbf{x}))$ (tai $P_{\mathbf{x}}(\mathbf{Y})$) (Shpitser ja Pearl, 2006b).

Interventiot ovat puhtaasti symbolisia eikä interventiota pystytä usein toteuttamaan konkreettisesti esimerkiksi suunnitellun kokeen avulla. Tällöin joudutaan turvautumaan havaittujen muuttujien jakaumaan $P(\mathbf{V})$. Kysymyksenä onkin tällöin, voidaanko kausaalivaikutus määrittää yksikäsitteisesti pelkästään jakauman $P(\mathbf{V})$ ja kausaalimallia M kuvaavan graafin G avulla. Kyse on *identifioituvuudesta*: kausaalivaikutus $P_{\mathbf{x}}(\mathbf{Y})$ on identifioituva mikäli $P^1_{\mathbf{x}}(\mathbf{Y}) = P^2_{\mathbf{x}}(\mathbf{Y})$ jokaiselle parille kausaalimalleja M^1 ja M^2, joille $P^1(\mathbf{V}) = P^2(\mathbf{V})$ ja joita vastaa sama graafi G.

Pearl (1995) kehitti joukon laskusääntöjä, jolla interventiojakaumia on mahdollista manipuloida identifioituvuuden määrittämiseksi. Tätä joukkoa kutsutaan *kausaalilaskennaksi* (*do-calculus*), jonka säännöt ovat:

1. Havaintojen lisääminen ja poistaminen

$$P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}, \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}),$$

   jos $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}}}}$ eli jos muuttujat $\mathbf{Y}$ ovat riippumattomia muuttujista $\mathbf{Z}$ ehdolla $\mathbf{X}$ ja $\mathbf{W}$ graafissa G, josta on poistettu solmujoukkoon $\mathbf{X}$ saapuvat särmät.

2. Toiminnan ja havainnon vaihtaminen

$$P_{\mathbf{x},\mathbf{z}}(\mathbf{y} \mid \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}, \mathbf{w}),$$

   jos $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}},\underline{\mathbf{Z}}}}$ eli jos muuttujat $\mathbf{Y}$ ovat riippumattomia muuttujista $\mathbf{Z}$ ehdolla $\mathbf{X}$ ja $\mathbf{W}$ graafissa G, josta on poistettu solmujoukkoon $\mathbf{X}$ saapuvat särmät ja solmujoukosta $\mathbf{Z}$ lähtevät särmät.

3. Toiminnan lisääminen ja poistaminen

$$P_{\mathbf{x},\mathbf{z}}(\mathbf{y} \mid \mathbf{w}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{w}),$$

jos $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}},\overline{Z(\mathbf{W})}}}$ eli jos muuttujat $\mathbf{Y}$ ovat riippumattomia muuttujista $\mathbf{Z}$ ehdolla $\mathbf{X}$ ja $\mathbf{W}$ graafissa G, josta on poistettu solmujoukkoon $\mathbf{X}$ saapuvat särmät ja solmujoukkoon $Z(\mathbf{W})$ saapuvat särmät, missä

$$Z(\mathbf{W}) = \mathbf{Z} \setminus An(\mathbf{W})_{G_{\overline{\mathbf{X}}}}$$

eli $Z(\mathbf{W})$ sisältää joukon $\mathbf{Z}$ ne solmut, jotka eivät kuulu joukkoon $\mathbf{W}$ ja eivät ole minkään joukon $\mathbf{W}$ solmun havaittuja esivanhempia graafissa G, josta on poistettu solmujoukkoon $\mathbf{X}$ saapuvat särmät.

Kausaalivaikutuksen identifioituvuus voidaan osoittaa, mikäli sääntöjä soveltamalla voidaan löytää kausaalivaikutukselle lauseke, jossa ei esiinny lainkaan do(·)-operaattoria eikä havaitsemattomia muuttujia. Säännöt itsessään eivät kuitenkaan kerro, missä järjestyksessä niitä tulisi soveltaa tai onko kiinnostuksen kohteena oleva kausaalivaikutus identifioituva.

Kausaalilaskennan algoritmi

Kausaalilaskentaan liittyvistä haasteista huolimatta on identifioituvuuden määrittämiseksi johdettu lukuisia tuloksia, joista tutkielmassa keskitytään ensisijaisesti Shpitserin ja Pearlin (2006) kehittämään kausaalilaskennan algoritmiin. Algoritmi kykenee määrittämään minkä tahansa kausaalivaikutuksen identifioituvuuden ja se tuottaa kausaalivaikutuksen lausekkeen yhteisjakauman $P(\mathbf{V})$ avulla esitettynä tapauksissa, joissa kausaalivaikutus on identifioituva. On syytä mainita, että myös Huang ja Valtorta (2006) johtivat vastaavan identifioituvuusalgoritmin samanaikaisesti. Molemmat algoritmit nojaavat vahvasti Tianin ja Pearlin (2003) aikaisempaan algoritmiin ja Tianin (2002) identifioituvuustuloksiin. Algoritmiin liittyviä teknisiä yksityiskohtia ja määritelmiä käsitellään tarkemmin tutkielmassa.

Algoritmin perusidea on hajottaa tarkasteltava graafi *c-komponentteihin*, jotka ovat sen maksimaalisia aligraafeja, joissa kaikkia solmupareja yhdistää pelkästään kaksisuuntaisista särmistä koostuva polku. Alkuperäinen ongelma voidaan tällöin hajottaa osa-ongelmiin, ja identifioituvuutta voidaan tarkastella jokaisessa c-komponentissa kerrallaan. Mikäli identifioituvuus ei toteudu jossain c-komponentissa, ei alkuperäiselläkään ongelmalla ole tällöin ratkaisua. Shpitser ja Pearl osoittivat tämän käyttäen kahdesta c-komponentista koostuvaa rakennetta, jota kutsutaan *pensasaidaksi* (*hedge*). Pensasaidan avulla on mahdollista esittää konstruktio, jossa kahdella kausaalimallilla on sama yhteisjakauma $P(\mathbf{V})$ ja graafi G, mutta ne tuottavat silti eri kausaalivaikutuksen, jolloin kausaalivaikutus ei voi olla identifioituva.

Tutkielmassa esitellään myös kausaalilaskennan algoritmin implementaatio R-kielellä, joka hyödyntää erityisesti R-paketteja *XML* (Temple Lang, 2013), *ggm* (Marchetti et al., 2015) ja *igraph* (Csardi ja Nepusz, 2006). Implementaatio on saatavilla CRAN-sivustolla pakettina *causaleffect* (https://cran.r-project.org/package=causaleffect). Pakettia on sittemmin laajennettu algoritmeilla muun muassa ehdollisten kausaalivaikutusten identifiointiin (Shpitser ja Pearl, 2006a), z-identifioituvuuden käsittelyyn (Bareinboim ja Pearl, 2012) ja kuljetettavuusongelman ratkaisuun (Bareinboim ja Pearl, 2013).

```
> library(causaleffect)
> library(igraph)

> g <- graph.formula(X -+ Y, Z_1 -+ X, Z_2 -+ Z_1,
 X -+ Z_2, Z_2 -+ X, Y -+ Z_2, Z_2 -+ Y, simplify = FALSE)
> g <- set.edge.attribute(g, "description", 4:7, "U")

> causal.effect(y = "Y", x = "X", G = g)
[1] "\\frac{\\sum_{Z_2}P(Y|Z_2,Z_1,X)P(X|Z_2,Z_1)P(Z_2)}
  {\\sum_{Z_2,Y}P(Y|Z_2,Z_1,X)P(X|Z_2,Z_1)P(Z_2)}"
```

Kuva 6.1.2: Esimerkki *causaleffect*-paketin käytöstä. Kaksisuuntaiset särmät määritellään kahtena yksisuuntaisena särmänä, joille on asetettava erityinen *description*-attribuutti erottamaan ne muista särmistä.

Esimerkiksi Kuvan 6.1.1 graafissa muuttujan X kausaalivaikutus muuttujaan Y voidaan identifioida *causaleffect*-paketilla käyttäen Kuvan 6.1.2 R-koodia. Kausaalivaikutukselle saadaan seuraava lauseke:

$$P_x(Y) = \frac{\sum_{Z_2} P(Y|Z_2, Z_1, X)P(X|Z_2, Z_1)P(Z_2)}{\sum_{Z_2,Y} P(Y|Z_2, Z_1, X)P(X|Z_2, Z_1)P(Z_2)}.$$

Kirjallisuus

[1] E. Bareinboim ja J. Pearl. “Causal inference by surrogate experiments: z-identifiability”. Teoksessa: *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*. Toim. N. de Freitas ja K. Murphy. AUAI Press, 2012, s. 113–120.

[2] E. Bareinboim ja J. Pearl. “Meta-Transportability of Causal Effects: A Formal Approach”. Teoksessa: *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*. 2013, s. 135–143.

[3] G. Csardi ja T. Nepusz. “The igraph software package for complex network research”. *InterJournal* Complex Systems (2006), s. 1695. URL: http://igraph.org.

[4] Y. Huang ja M. Valtorta. “Pearl’s calculus of intervention is complete”. Teoksessa: *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006, s. 217–224.

[5] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. New York: Guilford, 1998.

[6] D. Koller ja N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[7] Giovanni M. Marchetti, Mathias Drton ja Kayvan Sadeghi. *ggm: Functions for graphical Markov models*. R package version 2.3. 2015. URL: https://CRAN.R-project.org/package=ggm.

[8] J. Pearl. "Causal diagrams for empirical research". *Biometrika* 82 (4 1995), s. 669–688.

[9] J. Pearl. *Causality: Models, Reasoning and Inference*. 2nd. New York: Cambridge University Press, 2009.

[10] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". *Journal of Educational Psychology* 66 (5 1974), s. 688–701.

[11] I. Shpitser ja J. Pearl. "Identification of conditional interventional distributions". Teoksessa: *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006, s. 437–444.

[12] I. Shpitser ja J. Pearl. "Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models". Teoksessa: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*. Boston, Massachusetts: AAAI Press, 2006, s. 1219–1226.

[13] D. Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus*. R package version 3.98-1.1. 2013. URL: http://CRAN.R-project.org/package=XML.

[14] J. Tian ja J. Pearl. "A general identification condition for causal effects". Teoksessa: *Proceedings of the 18th National Conference on Artificial Intelligence*. AAAI/The MIT Press, 2002, s. 567–573.

[15] J. Tian ja J. Pearl. *On the identification of causal effects*. Tekninen raportti. R-290-L. Department of Computer Science, University of California, Los Angeles, 2003.

[16] T. S. Verma. *Graphical aspects of causal models*. Tekninen raportti. R-191. Department of Computer Science, University of California, Los Angeles, 1993.

Leo Törnqvist -palkinto

New approach to complex valued ICA: from FOBI to AMUSE

Niko Lietzén
Aalto University
School of Science
Department of Mathematics and
Systems Analysis

Abstract

A generic problem in different fields of science is to separate useful signals from noise and interferences. Statistical procedures that seek to reconstruct unobservable and interesting source signals, that are assumed to be mixtures of observable signals, are referred to as blind source separation (BSS) techniques. An important subclass of blind source separation is the independent component analysis (ICA), where we further assume that the interesting source signals are stochastically independent.

In the Master's Thesis "New Approach to Complex Valued ICA: From FOBI to AMUSE", we consider two famous blind source separation procedures for multivariate complex valued stochastic processes. Furthermore, we extend a performance measure called minimum distance index for complex valued settings. Defining performance measures for blind source separation procedures is not straightforward, since the model parameters are usually not uniquely defined. Additionally, we present simulation studies, that support the theory, and present some applications for complex valued blind source separation and independent component analysis.

Keywords: blind source separation, independent component analysis, complex valued statistics, performance measure, AMUSE, FOBI

The Thesis is available online: http://urn.fi/URN:NBN:fi:aalto-201604201851

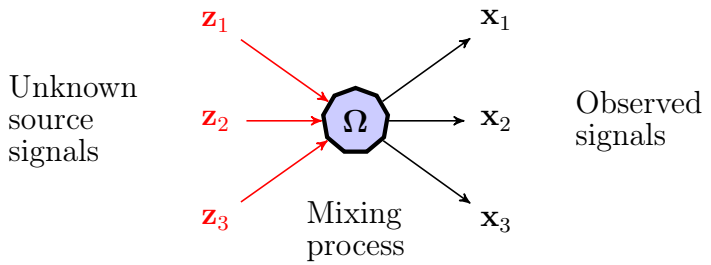


Figure 6.2.1: The blind source separation model.

Blind source separation

Blind source separation (BSS) is a class of statistical procedures that seek to recover source signals that have been mixed by some mixing system. The BSS problem is illustrated in Figure 6.2.1, where we have unknown signals, denoted by z_1, z_2, z_3, that enter an unknown mixing system $\mathbf{\Omega}$, such that the mixed signals, denoted by x_1, x_2, x_3, are the only observable part of the mixing process. An important subclass of blind source separation is the independent component analysis (ICA), where we further assume that the source signals are stochastically independent. The objectives in BSS usually are to find a transformation that reverses the effects of the mixing system $\mathbf{\Omega}$ and to recover the source signals.

The BSS problem has been considered in several different settings. The mixing system $\mathbf{\Omega}$ is typically assumed to be a linear transformation, although nonlinear transformations have also been studied especially in the case of independent component analysis, see Hyvärinen et al., 2001. The mixing system can also be defined such that some noise is added to the observed signals and the number of the source signals does not necessarily have to be equal to the number of observed signals. Furthermore, the BSS problem has been considered for source signals that are, e.g., real valued random variables, complex valued stochastic processes, tensor valued random variables, complex valued tensorial stochastic processes, see Ilmonen et al., 2010a; Lietzén et al., 2016; Virta et al., 2017; Lietzén et al., 2017.

In the Master's Thesis, we consider the following linear version of the BSS model. Let $\mathbf{x}_t := \{\mathbf{x}_t\}_{t\in\mathbb{N}}$ be a $\mathbb{C}^p$-valued stochastic process such that,

$$\mathbf{x}_t = \mathbf{\Omega}\mathbf{z}_t, \qquad \text{for every } t \in \mathbb{N}, \tag{6.2.1}$$

where $\mathbf{\Omega} \in \mathbb{C}^{p\times p}$ is invertible and the $\mathbb{C}^p$-valued stochastic process $\mathbf{z}_t$ satisfies,

$$\begin{aligned}
&\mathbb{E}[\mathbf{z}_t] = 0, && \text{for all } t \in \mathbb{N},\\
&\mathbb{E}\left[\mathbf{z}_t \mathbf{z}_t^{\mathrm{H}}\right] = 0, && \text{for all } t \in \mathbb{N},\\
&\mathbb{E}\left[\mathbf{z}_t \mathbf{z}_{t+\tau}^{\mathrm{H}}\right] = \mathbf{\Lambda}_\tau = \operatorname{diag}(\lambda_\tau^{(1)}, \ldots, \lambda_\tau^{(p)}), && \text{for all } t, \tau \in \mathbb{N},
\end{aligned}$$

and we denote the conjugate transpose of $\mathbf{X}$ as $\mathbf{X}^{\mathrm{H}}$. Under the assumptions, we have that $\mathbf{z}_t$ is weakly stationary. The goal is then to find a transformation matrix $\mathbf{\Gamma}$, such that the stochastic process $\{\mathbf{\Gamma}\mathbf{x}_t\}_{t\in\mathbb{N}}$ is diagonal with respect to the covariance matrix and the autocovariance matrix with lag τ. Note that in order to derive e.g. limiting distributions, we require some additional assumptions of

the existence of the moments and the regularity of the marginal distributions of $\mathbf{z}_t$.

The BSS model is not uniquely defined. Let $\tilde{\mathbf{z}}_t = \mathbf{PJDz}_t$ and $\tilde{\mathbf{\Omega}} = \mathbf{\Omega}(\mathbf{PJD}^{-1})^{\mathrm{H}}$, where $\mathbf{P}$ is any permutation matrix, $\mathbf{D}$ is any diagonal matrix such that the diagonal elements are in $\mathbb{R}_+$ and $\mathbf{J} = \mathrm{diag}(\exp(\theta_1 i), \ldots, \exp(\theta_p i))$ is any phase-shift matrix and i is the imaginary unit. Then $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{\Omega}}$ also satisfy Equation (6.2.1). This identifiability problem can be solved by either standardizing the components of $\mathbf{z}_t$ or by normalizing $\mathbf{\Omega}$, see Ilmonen et al., 2010a; Ilmonen, Paindaveine, et al., 2011.

BSS and ICA are applied in several fields of science, for a collection see Comon and Jutten, 2010. Complex valued BSS for stochastic processes are encountered especially in biomedical applications, where multiple Fourier and inverse Fourier transformations are implemented in the measurement process. For example, in functional magnetic resonance imaging the data can be complex valued and it has been recently discussed that relevant information is lost in the process if the complex valued structure is ignored, see Adali and Calhoun, 2007.

Performance measures

The main result of the Master's Thesis is the extension of the minimum distance (MD) index, introduced in Ilmonen et al., 2010b, for complex valued settings. Note that due to the identifiability problems of the BSS model (6.2.1), different BSS procedures do not necessarily estimate the same population quantities. The minimum distance index is a performance measure that ensures a fair comparison, even under these identifiability problems.

Let $\mathcal{C}$ denote the set of matrices of the form $\mathbf{DP}$, where $\mathbf{P}$ is a $p \times p$ permutation matrix and $\mathbf{D}$ is a complex valued $p \times p$ diagonal matrix. The sets $\{\mathbf{CA} : \mathbf{C} \in \mathcal{C}\}$ partition the set of complex valued $p \times p$ matrices into equivalence classes. If $\mathbf{B} \in \{\mathbf{CA} : \mathbf{C} \in \mathcal{C}\}$, notation $\mathbf{A} \sim \mathbf{B}$ is used. The shortest squared distance between the set $\{\mathbf{CA} : \mathbf{C} \in \mathcal{C}\}$, that is the set of matrices that are equivalent to $\mathbf{A}$, and $\mathbf{I}_p$ is given by

$$D^2(\mathbf{A}) = \frac{1}{p-1} \inf_{\mathbf{C} \in \mathcal{C}} ||\mathbf{CA} - \mathbf{I}_p||_{\mathrm{F}}^2, \tag{6.2.2}$$

where $||\cdot||_{\mathrm{F}}$ is the Frobenius norm.

Note that $D^2(\mathbf{A}) = D^2(\mathbf{CA})$ for all $\mathbf{C} \in \mathcal{C}$.

Theorem 6.2.1. *Let $\boldsymbol{A} \in \mathbb{C}^{p \times p}$ be any full rank matrix. The shortest squared distance $D^2(\boldsymbol{A})$ fulfills the following four conditions given below:*

1. $0 \le D^2(\boldsymbol{A}) \le 1$,
2. $D^2(\boldsymbol{A}) = 0$ *iff* $\boldsymbol{A} \sim \boldsymbol{I}_p$,
3. $D^2(\boldsymbol{A}) = 1$ *iff* $\boldsymbol{A} \sim \boldsymbol{1}_p \boldsymbol{a}^T$ *for some complex valued p-variate vector* $\boldsymbol{a}$, *and*
4. *the function* $c \mapsto D^2\left(I_p + c \cdot \mathrm{off}(\boldsymbol{A})\right)$ *is increasing in* $c \in [0,1]$ *for all matrices* $\boldsymbol{A}$ *such that* $|\boldsymbol{A}_{ij}| \le 1$, $i \ne j$.

Consider the complex valued BSS model with mixing matrix $\mathbf{\Omega}$ and an unmixing matrix estimate $\hat{\mathbf{\Gamma}}$. The shortest distance between the identity matrix and the set of matrices $\left\{\mathbf{C}\hat{\mathbf{G}} : \mathbf{C} \in \mathcal{C}\right\}$, that is matrices equivalent to the gain matrix $\hat{\mathbf{G}} = \hat{\mathbf{\Gamma}}\mathbf{\Omega}$, is given in the following definition.

Definition 6.2.2. The minimum distance index for the BSS unmixing estimate $\hat{\mathbf{\Gamma}}$ is

$$\hat{D} = D(\hat{\mathbf{\Gamma}}\mathbf{\Omega}) = \frac{1}{\sqrt{p-1}} \inf_{\mathbf{C}\in\mathcal{C}} \left\|\mathbf{C}\hat{\mathbf{\Gamma}}\mathbf{\Omega} - \mathbf{I}_p\right\|_{\mathrm{F}}.$$

From Theorem 6.2.1, we have that $0 \leq \hat{D} \leq 1$, and $\hat{D} = 0$ only if $\hat{\mathbf{\Gamma}} \sim \mathbf{\Omega}^{-1}$. Furthermore, $\hat{D} = 1$ is obtained in the pathological case when all the row vectors of $\hat{\mathbf{\Gamma}}\mathbf{\Omega}$ have the same direction. The value of the minimum distance index is now easy to interpret. Values close to 0 are associated with excellent separation, and large values indicate poor performance. Note that $D(\hat{\mathbf{\Gamma}}\mathbf{\Omega}) = D(\mathbf{C}\hat{\mathbf{\Gamma}}\mathbf{\Omega})$ for all $\mathbf{C} \in \mathcal{C}$.

The minimum distance index can then be implemented using the following theorem.

Theorem 6.2.3. *Let $\mathcal{P}$ denote the set of all $p \times p$ permutation matrices. Let $\hat{\boldsymbol{G}} = \hat{\mathbf{\Gamma}}\mathbf{\Omega}$ and let $\tilde{G}_{hj} = |\hat{G}|^2_{hj} / \sum_{k=1}^{p} |\hat{G}|^2_{hk}$. Now the minimum distance index can be expressed as*

$$\hat{D} = D(\hat{\boldsymbol{G}}) = \frac{1}{\sqrt{p-1}} \left(p - \max_{\boldsymbol{P}\in\mathcal{P}} \left(\operatorname{tr}\left(\boldsymbol{P}\tilde{\boldsymbol{G}}\right)\right)\right)^{1/2}.$$

The maximization problem

$$\max_{\mathbf{P}\in\mathcal{P}} \left(\operatorname{tr}(\mathbf{P}\tilde{\mathbf{G}})\right) \tag{6.2.3}$$

over all permutation matrices P is equivalent to the optimization problem called linear sum assignment problem (LSPA). In our approach, we solve the LSPA using the Hungarian method, see Papadimitriou and Steiglitz (1982).

Bibliography

[1] Tülay Adali and Vince D Calhoun. "Complex ICA of brain imaging data". In: *IEEE Signal Processing Magazine* 24.5 (2007), p. 136.

[2] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[3] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[4] Pauliina Ilmonen, Jaakko Nevalainen, and Hannu Oja. "Characteristics of multivariate distributions and the invariant coordinate system". In: *Statistics & probability letters* 80.23-24 (2010), pp. 1844–1853.

[5] Pauliina Ilmonen, Klaus Nordhausen, Hannu Oja, and Esa Ollila. "A new performance index for ICA: properties, computation and asymptotic analysis". In: *Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 229–236.

[6] Pauliina Ilmonen, Davy Paindaveine, et al. “Semiparametrically efficient inference based on signed ranks in symmetric independent component models”. In: *the Annals of Statistics* 39.5 (2011), pp. 2448–2476.

[7] Niko Lietzén, Klaus Nordhausen, and Pauliina Ilmonen. “Complex Valued Robust Multidimensional SOBI”. In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer. 2017, pp. 131–140.

[8] Niko Lietzén, Klaus Nordhausen, and Pauliina Ilmonen. “Minimum distance index for complex valued ICA”. In: *Statistics & Probability Letters* 118 (2016), pp. 100–106.

[9] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1982.

[10] Joni Virta, Bing Li, Klaus Nordhausen, and Hannu Oja. “Independent component analysis for tensor-valued data”. In: *Journal of Multivariate Analysis* 162 (2017), pp. 172–192.

Päätäntätiedettä Suomessa 1968 ja 2018

Juha Karvanen
Jyväskylän yliopisto

Tiivistelmä

Päätösanalytiikkaa sovelletaan nykyään monella eri alalla. Suomessa päätäntätieteestä kirjoittivat jo vuonna 1968 Leo Törnqvist ja Leif Nordberg.

Viisikymmentä vuotta sitten Leo Törnqvist ja Leif Nordberg julkaisivat kirjan ”Päätäntätieteen keskeisiä ongelmia” (Törnqvist ja Nordberg, 1968). Törnqvist (1911–1983) oli Helsingin yliopiston ensimmäinen tilastotieteen professori, jonka mukaan Tilastoseuran jakama pro gradu -palkinto on nimetty. Nordberg (1943–) toimi tilastotieteen ja ekonometrian professorina Åbo Akademissa vuosina 1975–2007 (Professoriliitto, 2008). Kirjassa on vain 64 sivua, mutta sen tavoite on silti kunnianhimoinen: luoda systemaattinen perusta päätäntätieteelle.

Törnqvist ja Nordberg määrittelevät peruskäsitteistön, johon kuuluvat esimerkiksi termit päätöstentekijä, päätöstilanne, päättämisprosessi, toteuttamisprosessi ja päätäntätoiminta. Matemaattisen esityksen lisäksi tekijät pohtivat päätöksentekoa yhteiskunnallisesta ja inhimillisestä näkökulmasta ja esittävät muun muassa, että päätöksentekijällä tulee olla ”hyvänsuopa ja leppoisa asenne toisia päätöksentekijöitä kohtaan” (Törnqvist ja Nordberg, 1968, s. 61).

Törnqvistin ja Nordbergin kirjan julkaisemisen aikaan päätöksentekoon liittyvät kysymykset olivat kansainvälisen huomion kohteena. Joitakin vuosia aiemmin ilmestynyt Raiffan ja Schlaiferin (1961) kirja on vielä nykyäänkin usein käytetty viite. Päätöksentekoteorian keskeinen periaate on hyötyfunktion odotusarvon maksimointi. Laskennallisesti tämä oli kuitenkin 1960-luvulla ja vielä pitkään sen jälkeenkin liian vaativaa teorian täysimittainen soveltamisen kannalta. Hyötyfunktion täsmällinen määrittäminenkin on usein päätöksentekijälle vaikeaa, etenkin, jos tilanteeseen liittyy useita ristiriitaisia tavoitteita.

Vuonna 2018 päätäntätiedettä, tai nykyaikaisemmin päätösanalytiikkaa, sovelletaan Suomessa monella eri alalla. Esimerkiksi Jyväskylän yliopisto on valinnut alan yhdeksi profiloitumiskohteistaan otsikolla ”Decision analytics utilising causal models and multiobjective optimization”. Tavoitteena on tilastotiedettä, koneoppimista ja optimointia käyttäen rakentaa saumaton ketju datasta päätökseen. Interaktiiviset menetelmät auttavat päätöksentekijää ilmaisemaan preferenssinsa monitavoitteisessa päätöksenteossa (Miettinen, 2014).

Suomalaiset nykytilastotieteilijät ovat soveltaneet päätöksentekoteoriaa esimerkiksi metsätieteissä (Alho ja Kangas, 1997; Islam et al., 2010) ja kalavesien hoidossa (Kuikka et al., 2014; Mäntyniemi et al., 2009). Nämä sovellukset eivät olisi olleet mahdollisia ilman laskentatehon kasvua ja Bayes-menetelmien kehitystä. Otanta ja koesuunnittelu voidaan myös ymmärtää päätösongelmiksi. Tutkimusasetelman optimointia ovat Suomessa käsitelleet muun muassa Tokola et al. (2014), Mehtälä et al. (2015) ja Reinikainen et al. (2016). Voitaneen sanoa, että alalla on suomalaisen tilastotieteen kokonaisvolyymiin suhteutettuna merkittävää kotimaista tutkimustoimintaa. Ehkäpä olemme kulkemassa Törnqvistin ja Nordbergin viimeisen luvun otsikon mukaisesti kohti parempaa päätäntätoimintaa.

Kirjallisuus

[1] Juha M Alho ja Jyrki Kangas. "Analyzing uncertainties in experts' opinions of forest plan performance". *Forest Science* 43.4 (1997), s. 521–528.

[2] Md Nurul Islam, Mikko Kurttila, Lauri Mehtätalo ja Timo Pukkala. "Inoptimality losses in forest management decisions caused by errors in an inventory based on airborne laser scanning and aerial photographs". *Canadian Journal of Forest Research* 40.12 (2010), s. 2427–2438.

[3] Sakari Kuikka, Jarno Vanhatalo, Henni Pulkkinen, Samu Mäntyniemi, Jukka Corander et al. "Experiences in Bayesian inference in Baltic salmon management". *Statistical Science* 29.1 (2014), s. 42–49.

[4] Juha Mehtälä, Kari Auranen ja Sangita Kulathinal. "Optimal observation times for multistate Markov models–applications to pneumococcal colonization studies". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.3 (2015), s. 451–468.

[5] Samu Mäntyniemi, Sakari Kuikka, Mika Rahikainen, Laurence T. Kell ja Veijo Kaitala. "The value of information in fisheries management: North Sea herring as an example". *ICES Journal of Marine Science* 66.10 (2009), s. 2278–2283.

[6] Kaisa Miettinen. "Survey of methods to visualize alternatives in multiple criteria decision making problems". *OR spectrum* 36.1 (2014), s. 3–37.

[7] Professoriliitto. *Suomen professorit 1640–2007*. toim. L. Ellonen. Helsinki: Professoriliitto, 2008.

[8] Howard Raiffa ja Robert Schlaifer. *Applied Statistical Decision Theory*. Boston: Harvard University, 1961.

[9] Jaakko Reinikainen, Juha Karvanen ja Hanna Tolonen. "Optimal selection of individuals for repeated covariate measurements in follow-up studies". *Statistical Methods in Medical Research* 25.6 (2016), s. 2420–2433.

[10] L. Törnqvist ja L. Nordberg. *Päätäntätieteen keskeisiä ongelmia*. Porvoo: WSOY, 1968.

[11] Kari Tokola, Andreas Lundell, Jaakko Nevalainen ja Hannu Oja. “Design and cost optimization for hierarchical data”. *Statistica Neerlandica* 68.2 (2014), s. 130–148.

Finnish young statisticians workshop 2018

Paavo Raittinen

The Finnish Statistical Society, The Finnish Society of Biostatistics, and Aalto University School of Science organized a workshop for young Finnish statisticians on 20th of November 2018. The event was named Finnish Young Statisticians Workshop 2018 – FYSW2018. The original motivation behind the event was to bring all the doctoral students from different subfields of statistics and encourage them to give presentations on an easy-to-approach environment.

Each participant was given an opportunity to give a 25 minute presentation. In total, FYSW2018 had 12 participants, and six presentations. Topics spanned from joint modeling in an epidemiologic setting to asymptotic properties of a blind source separation estimator. There was something new, and surely something inspiring, for everyone. Some interesting statistics from FYSW2018: the participation rate from University of Eastern Finland and University of Turku was 100%, that is, all statistics doctoral students were present from UEF and UTU. Furthermore, in terms of academic ripeness distribution, there were participants from those who had just begun their doctoral studies to those who had already started asking for a bookbindery contacts. Overall, the event was a huge success and the social event after the talks allowed more easy-going discussion and exchange of ideas. The feedback sounded uniform: the FYSW was worth going to, and definitely worth going again next year. Hopefully this will lead to an annual gathering of young statisticians and helps to tighten their network.

Kesässä kandiksi

Petteri Mäntymaa
Terveyden ja Hyvinvoinnin Laitos THL

Tilastotieteen toisen opiskeluvuoteni saapuessa päätökseen sain ensimmäisen oman alani työpaikan korkeakouluharjoittelijana Tilastokeskuksella. Koskaan ei ollut uuden työn aloittaminen ollut näin jännittävää. Epävarmuus omasta osaamisesta oli valtava, tuntuivathan opinnot olevan vasta aluillaan, ja kandidaatin tutkielmakin tuntui häämöttävän vasta jossain kaukaisessa tulevaisuudessa.

Tilastokeskus työllistää kesäisin laajan joukon opiskelijoita eri puolilta Suomea. Toisin kuin voisi kuvitella, korkeakouluharjoittelijat muodostavat opintosuunniltaan ja koulutusohjelmiltaan hyvin monitieteellisen joukon, aina kansatieteen opiskelijoista tuleviin sosiologeihin. Toki mukaan mahtuu niin muutama talous- ja tilastotieteilijä kuin matemaatikkokin.

Oma tehtäväni oli empiirinen selvitystyö Euroopan komission rahoittamasta ja Hollannin tilastoviraston koordinoimasta *Representativity indicators for survey quality*, eli RISQ-projektista. Tavoitteena oli tehdä katsaus erään tällaisen kyselytutkimusaineiston edustavuusmittarin, *R-indikaattorin* teoriaan sekä tutkia sen soveltuvuutta Tilastokeskuksen kyselytutkimusaineistoihin.

Vastauskato on alati kasvava ongelma kyselytutkimuksissa ja sen vaikutus tutkimusten laatuun lienee kiistämätön. Kun otokseen valikoituneilta ei saada tarvittavaa informaatiota, vastanneiden joukon koko pienenee ja perusjoukon estimaattien epävarmuus kasvaa. Vastausastetta käytetäänkin usein kyselytutkimuksen ensisijaisena laatuindikaattorina. R-indikaattoriin liittyvässä kirjallisuudessa erityinen huoli liittyy kuitenkin vastauskäyttäytymisen heterogeenisyyteen. Mikäli vastaamattomuus korostuu ryhmissä, joiden vastaukset mahdollisesti poikkeaisivat muista ryhmistä, myös estimaattien harha kasvaa. R-indikaattorin teoria perustuukin aineiston *edustavuuden* arviointiin, eli siihen kuinka paljon eri ryhmien estimoidut vastaustodennäköisyydet poikkeavat toisistaan.

Mukavan kokoista artikkelikokoelmaa läpi kahlatessani aloin työstämään tiivistelmää R-indikaattorin teoriasta ja samalla testaamaan millaisia tuloksia olisi R-indikaattorin avulla saatavissa Tilastokeskuksen Työvoimatutkimuksesta, Väestön tieto- ja viestintäteknologian käyttö -tutkimuksesta sekä European Social Surveystä.

Tilastokeskus osoittautui hyvin antoisaksi paikaksi saada ensikosketus tilastotieteilijän työhön, sillä talossa ja etenkin omassa tiimissäni oli todella hyvin määritellyt toimintatavat korkeakouluharjoittelijoiden perehdyttämiseen sekä työnohjaukseen. Pääsin jo kesän alkuvaiheessa esittelemään työtäni niin tiimin sisäisissä ja myöhemmin myös ulkoisissa tilastoseminaareissa. Palaute oli todella rakentavaa, kannustavaa. Tehokkaassa ohjauksessa työtkin etenivät su-

juvasti. Myönnettäköön, että tarkkaa hetkeä tai paikkaa en muista, mutta langanpäiden hiljalleen kohdatessa, ehkä myös jonkun tilannetta vierestä seuranneen kommentin kirvoittamana tajusin, että tässähän se kandidaatin tutkielma taitaa puolivahingossa tulla tehdyksi.

Omien kokemusteni mukaan kandidaatin tutkielma voi osalle opiskelijoista vielä kolmannenkin opiskeluvuoden alkaessa tuntua hyvin kaukaiselta ja helposti hamaan tulevaisuuteen lykättävältä. Tätä kynnystä on toki yliopiston taholta koitettu aktiivisesti madaltaa, mutta kaikki lisäponsi on varmasti tervetullutta. Toisaalta myös kynnys hakea oman alan töitä on korkea. Eikä ihme, matematiikka, jos jokin, opettaa nöyräksi. Graduharjoittelu on jo monella alalla vakiintunut käytäntö ja yhteistyö on parhaimmillaan molempia osapuolia hyödyttävää. Kandidaatin tutkielmaan tähtäävästä harjoittelusta voisi syntyä samankaltaisia hyötyjä. Opiskelijat saisivat matalan kynnyksen sisäänajon työelämään, valmistumisajat lyhenisivät ja työnantajat saisivat potentiaalisia tulevaisuuden työntekijäehdokkaita. Jotkin alan toimijat tiettävästi jo pohtivatkin tämän mahdollisuuksia, mutta toimintatavan soisi vakiintuvan ja leviävän laajemmallekin.

Afternoon seminar — Non-participation in population surveys

30 August 2017
National Institute for Health and Welfare THL
Mannerheimintie 166 (main building), 00300 Helsinki

Programme

1st session

Chair: Juha Karvanen, University of Jyväskylä

12:30-12:40 Welcome and introduction (Juha Karvanen)

12:40-13:00 Non-participation in health examination surveys - why is this a problem? (Kari Kuulasmaa)

13:00-13:20 Socio-economic position as a determinant of non-response (Jaakko Reinikainen)

13:20-13:40 Survey non-participants have worse health and health behaviours than participants (Pekka Jousilahti)

13:40-14:00 Ways to tackle with non-participation during the implementation of the survey (Hanna Tolonen)

2nd session

Chair: Kari Kuulasmaa, National Institute for Health and Welfare (THL)

14:30-14:50 Using re-contact or non-response questionnaire data to adjust for selective non-participation (Juha Karvanen)

14:50-15:10 Using register data to adjust for selective non-participation in cross-sectional survey setting (Juho Kopra)

15:10-15:30 Age-period-cohort mortality analysis of long-term differences between survey participants and nonparticipants (Tommi Härkänen)

Panel Discussion

Chair: Hanna Tolonen

15:30-15:40 Introduction to the panel discussion (Hanna Tolonen)

15:40-16:10 Panel Discussion

Kari Djerf (Tilastokeskus)

Päivikki Koponen(THL)

Jari Pajunen (Taloustutkimus)

Oona Pentala-Nikulainen (THL)

Reijo Sund (Helsingin yliopisto, Itä-Suomen yliopisto)

16:10-16:20 Conclusions (Hanna Tolonen)

The seminar was organized in collaboration with National Institute for Health and Welfare, and the Department of Mathematics and Statistics at the University of Jyväskylä.

Valikoituneen osallistujakadon tilastollinen mallintaminen terveystarkastustutkimuksessa

Juho Kopra
Jyväskylän yliopisto

Terveystarkastustutkimusten tavoitteena on kerätä luotettavaa tietoa kohdepopulaation terveydentilasta ja riskitekijöistä. Eräs keskeisimmistä luotettavuuteen vaikuttavista tekijöistä on puuttuva tieto. Terveystarkastustutkimuksissa puuttuvaa tietoa syntyy, kun osa tutkimukseen kutsutuista ei osallistu tutkimukseen, jolloin puhutaan poisjääneistä ja osallistujakadosta tai vain kadosta. Mikäli poisjäänti on yhteydessä tutkittaviin terveydellisiin tekijöihin, niin tutkimuksen osallistujilta lasketut tulokset eivät ole yleistettävissä alkuperäiseen kohdepopulaatioon eli luotettavuus heikkenee. Tällöin sanotaan, että osallistujien tiedoista lasketuissa estimaateissa on valikoitumisharhaa.

Väitöstutkimuksessani (Kopra, 2018) oli tavoitteena kehittää menetelmiä, joiden avulla voidaan pienentää valikoitumisharhaa poikkileikkausaineistossa lisätietoaineistoja käyttämällä. Käytössä oli aineisto kansallisesta FINRISKI-tutkimuksesta ja kiinnostuksen kohteena oli itseraportoidun päivittäisen tupakoinnin ja alkoholin suurkulutuksen vallitsevuus eli prevalenssi. Aiempien tutkimusten perusteella voitiin ennakolta arvioida että valikoitumisharhaa on, mutta ei kyetty sanomaan kuinka paljon. Lisätiedon lähteinä olivat 1) rekisteriperäiset sairaalakäynti- ja kuolinsyyaineistot (ns. seuranta-aineisto) sekä 2) nk. uudelleenyhteydenottoaineisto, joka oli kerätty terveystarkastuksesta poisjääneiltä ottamalla uudelleen yhteyttä terveystarkastuksen jälkeen. Seuranta-aineiston avulla voidaan saada epäsuoraa informaatiota kohdepopulaation tutkimushetken terveydentilasta ja elintavoista, mutta uuden yhteydenoton kautta saadaan vastaavaa tietoa kuin varsinaisessa terveystarkastustutkimuksessa pian kyselyn jälkeen. Väitöskirjassani esitän menetelmiä kumpaankin tilanteeseen.

Jos tupakoitsijat tai alkoholin suurkuluttajat osallistuvat harvemmin kuin muut, niin tupakoinnin yleisyyden estimaatit ovat alaspäin harhaiset. Tästä kertovat ei-osallistujien suurempi riski sairastua tai kuolla tupakointiin yhteydessä oleviin tauteihin, sekä aineistot, joissa ei-osallistuneita on pyydetty uudelleen vastaamaan kyselyyn. Ongelmaa vakavoittaa se, että nykyisin tyypillisesti vain 50–60% tutkimukseen kutsutuista osallistuu, kun vielä 1970-luvulla yli 90% osallistui. Mitä suurempi kato, sitä suurempi voi olla myös harha.

Uudelleenyhteydenottoaineistoa käytettäessä sovelsin moni-imputointia ja seuranta-aineistoa käytettäessä bayeslaista tilastollista mallintamista. Moni-imputointi FINRISKI-aineistolle perustui siihen että ei-osallistuneita kysyttiin uudelleen vastaamaan kyselyyn, jolloin saatiin tietoa heidän tupakoinnista ja alkoholinkäytöstään. Näin saatiin otos, joka ei välttämättä ole satunnaisotos. Seuranta-aineistoa hyväksikäyttäen huomasin, että saatu otos kuvasi varsin hyvin myös niitä, joilta ei saatu vastausta alkuperäiseen terveystarkastuskyselyyn eikä uudelleenyhteydenottokierrokseenkaan. Tämä voitiin päätellä siitä, että ei-osallistuneiden sairastuneisuus ja kuolleisuus oli samanlaista kuin uudelleenyhteydenottoon vastanneiden. Moni-imputointimallissa käytin oletusta, että uudelleenyhteydenottokierrokseen vastanneet ovat satunnaisotos terveystarkastuksen ei-osallistuneista ehdolla taustatiedot. Seuranta-aineistoon perustuen kyseinen oletus sai eniten tukea aineistolta verrattuna rajoittavampiin oletuksiin.

Bayes-mallinnuksessa keskeisenä ajatuksena oli hyödyntää tupakoinnin ja keuhkoahtaumataudin sekä keuhkosyövän tunnettua vahvaa yhteyttä aineistossa. Tautitapahtumat siis antavat lisäinformaatiota tupakoinnin yleisyydestä aineistossa. Vastaavasti alkoholin suurkulutuksen osalta käytettiin tauteja, jotka ovat vahvoja indikaattoreita alkoholin suurkulutuksesta. Käytetty Bayes-malli koostui kolmesta osamallista sekä tupakoinnin että alkoholin mallinnuksessa, joille kummallekin toteutettiin omat aivan erilliset mallinsa. Tautiriskiä (osamalli 1) mallinnettiin itseraportoidulla tupakoinnilla tai alkoholin suurkulutuksella sekä taustatiedoilla kuten ikä, sukupuoli ja asuinmaakunta Suomessa. Tupakointia ja alkoholin suurkulutusta (osamalli 2) mallinnettiin taustatiedoilla ja osallistumista tutkimukseen (osamalli 3) mallinnettiin sekä taustatiedoilla että tupakoinnilla tai alkoholin suurkulutuksella. Viimeksi mainittu osamalli salli alkoholin tai tupakoinnin vaikuttaa osallistumiseen, eli kyseinen malli olettaa puuttuvuuden olevan ei-satunnaista (Missing Not at Random, MNAR). Puuttuvuuden valikoituvuuden astetta määrittävälle parametrille asetettiin informatiivinen priori, mutta muut käytetyt priorijakaumat olivat epäinformatiivisia. Puuttuvien tiedot imputoitiin samalla kun Bayes-malli sovitettiin aineistoon. Informaatiota puuttuvien tietojen paikkaamiseen saatiin paitsi taustatiedoista niin varsinkin seuranta-aineistosta. Mallin sovittaminen oli laskennallisesti hyvin raskasta, koska aineistossa oli noin 10000 puuttuvaa tupakointi- tai alkoholitietoa, jotka simuloitiin jokaisella MCMC-algoritmin iteraatiolla. Jokaisen kokeillun mallin sovittaminen kesti useita päiviä ja mallin kehitystyö oli tästä johtuen vaikeaa.

Tällaista Bayes-mallinnusta hankaloittaa se, että seuranta-aineisto tarvitaan sekä osallistuneille että myös ei-osallistuneille. Voi kulua vuosia tai vuosikymmeniä, ennen kuin tupakointiin tai alkoholin suurkulutukseen liittyviä tautitapahtumia on tarpeeksi. Tästä syystä seuranta-aineiston käyttö ei ole käyttökelpoinen valikoitumisharhan korjaamiseen heti terveystarkastusaineiston keräämisen jälkeen.

Moni-imputointimallilla ja uudelleenyhteydenottoaineistolla selvisi, että päivittäisen tupakoinnin ja alkoholin suurkulutuksen yleisyydet olivat 2-4 prosenttiyksikköä korkeammat kuin mitä havaittiin pelkästään terveystarkastukseen osallistujilta (1.2-1.5 kertaa isompi). Nuorten miesten kohdalla alkoholin suurkulutuksen osuudeksi estimoitiin jopa 15.9% (95% luottamusväli 12.5-19.4), mikä oli korkeampi kuin muissa demografisissa ikä- ja sukupuoliryhmissä. Vuoden 2012 aineistosta, joka oli viimeisin käytössä ollut, estimoitiin tupakoinnin yleisyydeksi 28.5% miehille ja 19.0% naisille.

Saaduista tutkimustuloksista nähtiin, että lisätietoaineistoja ja käytettyjä ti-

lastomenetelmiä hyödyntämällä saatiin korkeammat vallitsevuusestimaatit päivittäiselle tupakoinnille ja alkoholin suurkulutukselle kuin perustuen pelkästään terveystarkastukseen osallistujilta saatuun aineistoon. Lisäksi vallitsevuusestimaattien luotto- tai luottamusvälit olivat leveämmät kuin pelkästään osallistujilta laskemalla olisi saatu. Osallistujien aineistosta lasketut kapeammat luotto- ja luottamusvälit perustuivat olettamukseen että valikoitumista ei ollut, mikä ei ole pätevä oletus tässä aineistossa. Tähän perustuen voidaan sanoa että leveämmät luotto- tai luottamusvälit ovat realistisempia. Moni-imputointia voidaan käyttää apuna vallitsevuuden harhan pienentämisessä, mikäli uudelleenyhteydenottoaineisto on kerätty. Bayeslainen mallintaminen soveltuu erityisesti tilanteeseen, jossa uudelleenyhteydenottoaineistoa ei ole saatavilla, mutta seuranta-aineistoa riittävän pitkällä seuranta-ajalla voidaan hyödyntää, jolloin seuranta-aineistosta siis saadaan riittävästi epäsuoraa tietoa osallistujien ja poisjääneiden terveydentilasta ja terveyskäyttäytymisestä.

Tulevissa terveystarkastustutkimuksissa kannattaa kerätä uudelleenyhteydenottoaineistoa ja tutkia myös vaihtoehtoisia mallinnusolettamuksia hyödyntäen kerättyihin datoihin yhdistettyjä seuranta-aineistoja. Potentiaalisia aiheita jatkotutkimukselle ovat mallien estimoinnin laskennallisen tehokkuuden parantaminen ja yhdenaikainen uudelleenyhteydenottoaineston ja seuranta-aineiston käyttö tilastollisessa mallinnuksessa.

Aiemmat tulokset perustuen pelkästään osallistuneilta kerättyyn dataan näyttävät olevan harhaisia, arvioivat epävarmuuden liian pieneksi, ja antavat liian positiivisen kuvan päivittäisen tupakoinnin ja alkoholin suurkulutuksen yleisyydestä. Työssä käytettyjä tilastomenetelmiä voidaan soveltaa myös muilla tieteenaloilla kuin terveystieteissä.

Kirjallisuus

[1] Juho Kopra. “Statistical modelling of selective non-participation in health examination surveys (Väitöskirja)”. *Report/University of Jyväskylä, Department of Mathematics and Statistics* (2018).

Afternoon seminar — Statistics in law

29 January 2018
Juhlasali, Language Centre (Kielikeskus), University of Helsinki
Fabianinkatu 26, Helsinki

Programme

1st session

12:00 Opening of the Seminar

12:10 Social disadvantage and crime in Finland: contrasting results from between-individual and within-individual models (Mikko Aaltonen)

12:50 Sentences and prosecutors' demands for aggravated drunk driving in Finland (Pekka Pere)

13:30 Multilevel modelling of sentencing: Variation between drunk driving in Finland (Mika Sutela)

2nd session

15:00 Natural selection': Conceptual issues and empirical strategies for treating sample selection bias in sentencing research (Brian Johnson)

15:40 Exploring disparities in sentencing using multilevel modelling: opportunities and pitfalls (Jose Pina-Sanchez)

16.20 Closing of the Seminar

The seminar was organized in collaboration with the Law School at the University of Eastern Finland.

Social disadvantage and crime in Finland: contrasting results from between-individual and within-individual models

Mikko Aaltonen Finnish Ministry of Justice

Finnish criminal policy has traditionally relied on the notion that social disadvantage causes crime, and thus prevention of social exclusion should be one of the focal points of effective crime prevention. Analyses of socioeconomic backgrounds of convicted offenders appear to confirm this hypothesis: offenders often lack educational qualifications and have weak ties to the labor market, leading to low incomes and debt problems. However, for a variety of reasons, the causal nature of these associations has proven difficult to establish. In this presentation, I present key results from a series of published and ongoing articles that have used individual-level register-based panel data and attempted to overcome some selection issues inherent to the study of these topics. Overall, these studies suggest that within-individual change in employment and social disadvantage tend to predict changes in property crime, but not violent crime. I conclude by discussing some recent studies that have dealt with selection bias in a convincing manner, and whether such studies could be feasible in Finland.

Sentences and prosecutors' demands for aggravated drunk driving in Finland

Pekka Pere University of Helsinki
Tuomas Lahti Independent scholar
Mika Sutela University of Eastern Finland
http://www.tandfonline.com/eprint/Gfb8pR5UssHWQS4SaWht/full

Sentences and prosecutors' demands for aggravated drunk driving are categorised into three classes: The sentence is more lenient than, is compatible with, or is harsher than the prosecutor's demand. The probability of a sentence falling into one of the three ordered categories is explained by a cumulative logit model. The following circumstances affect the probability of a more lenient or harsher sentence, in decreasing order of importance: driving a truck, facing at least four counts, having a legal assistant, and being present in the trial. The hypothesis that factors known by the prosecutor, at the time of writing the demand, should not systematically affect sentences is refuted. The judges assess circumstances differently than the prosecutors. The prosecutors' role is nevertheless prominent in the sense that the sentences follow, to a great extent, their demands. Notable gender effects of the actors in the courtroom are found.

Multilevel modelling of sentencing: Variation between individual judges and prosecutors in the severerity of punishments

Mika Sutela University of Eastern Finland

According to section 6 of the Constitution of Finland, everyone is equal before the law. In sentencing equality means that, in principle, the same criminal act should be sentenced by the same punishment regardless of where the act has been committed and who gives the judgment. In this multilevel study, criminal sentencing decisions is analyzed with linear mixed models. In particular, the variation in the severity of punishments between individual judges and prosecutors is examined. The research data consists of the aggravated drunk driving (1.20- ‰) cases (N = 477) in Finnish district courts during years 2006–2010. With the mixed models it is possible to take into account the hierarchical and nested structure of sentencing data. For instance, individual judges or prosecutors are nested with courts, and criminal cases are nested with individual judges and prosecutors. In the analysis, judges and prosecutors are handled as random effects, as well as district courts and years. Individual characteristics of courtroom actors and procedural/organizational/community-level factors are handled as fixed effects. Preliminary results show that legal factors are significant but also the variation between prosecutors is rather large.

'Natural selection': Conceptual issues and empirical strategies for treating sample selection bias in sentencing research

Brian Johnson University of Maryland

Analytical issues involving sample selection are pervasive in sentencing research. This lecture will provide a basic conceptual overview of sample selection bias in criminological research, along with an introductory treatment of common statistical modeling approaches that are designed to deal with different types of sample selection. It will focus primarily on the use of Heckman's selection model and its commonly used alternatives in the context of sentencing research, and it will provide a brief discussion of ongoing and emergent issues in this area.

Exploring disparities in sentencing using multilevel modelling: opportunities and pitfalls

Jose Pina-Sánchez University of Leeds

Unwarranted judicial disparities undermine trust in the criminal justice system. To tackle this problem a growing number of jurisdictions have followed the example of the US and implemented guidelines schemes seeking to constrain judicial discretion and promote consistency in sentencing. In this talk I will present the many opportunities that multilevel modelling techniques afford us to

explore the nature and extent of sentencing disparities that cannot be explained by legitimate case characteristics; how findings from such models can be used to rethink sentencing guidelines; but also how the underlying assumptions made by multilevel models need to be carefully considered.

More technically, I will present a variety of multilevel models that I have implemented over the last four years to explore the topic of unwarranted disparities in sentencing. Ranging from the standard random intercepts model (used to explore the share of unobserved variability in sentence length due to systematic disparities between courts) to the more complex location-scale model (used to estimate different levels of unexplained within court variability), and cross-classified models (capable of distinguishing unobserved variability stemming from the judge level, the court level, and the interaction between them).

Iltapäiväseminaari — Mikä on luotettavaa tietoa ja mistä se tunnistetaan? Virallinen tilasto muuttuvilla markkinoilla

12. Huhtikuuta 2018
Laituri, Narinkka 2, 00100 Helsinki

Ohjelma

13:00 Avaus (Jyrki Möttönen, Tilastoseura)

13:10 **Tiedon käyttäjät**

Tiedon hyödyntämisestä tiedolla johtamiseen (Jenni Airaksinen, Kuntaliitto)

Kaupunkitilasto 2020 — kaupunkitilaston kasvava kysyntä ja uudet haasteet (Ari Jaakola, Helsingin kaupunki

14:00 Kahvitauko

14:20 **Tilaston tuottajat ja välittäjät**

Tilasto ymmärryksen lisääjänä (Timo Koskimäki, Tilastokeskus)

15:00 **Tilastotieteen osaajat**

Tilastotieteilijä 2060-luvulla (Maria Valaste, Helsingin yliopisto)

15:20 Kokemuksia EMOS-ohjelmasta (Xiaoyuan Li, Helsingin yliopisto)

15:40 Seminaarin päätös (Pauliina Ilmonen, Tilastoseura)

Tilastot muuttuvassa maailmassa

Asta Manninen
EU:n ESAC-komitea
(European Statistical Advisory Committee)

Tilastojen toimintaympäristön muutoksia

Tietomarkkinat tänä päivänä ovat moninaiset. Tietoa on valtavasti tarjolla ja yhä nopeatempoisemmin. On monia erilaisia toimijoita ja palveluita.

Tilaston tuottamisen toimintaympäristö on muuttunut. Muutokseen ovat vaikuttaneet monet asiat, kuten digitalisaatio ja uudet aineistot (Big Data, IoT, citizen science ym.) ja globalisaatio. Tilaston julkaisemisen ja välittämisen tavat ja välineet ovat nekin digitalisaation myötä muuttuneet. Lisäksi niukkenevat budjetit haastavat virallisen tilaston laatijoita. Tässä toimintaympäristössä virallisen tilaston velvoite on jatkuvasti uudistua ja vastata näin käyttäjien ja muuttuvan yhteiskunnan tietotarpeisiin. Onneksi virallisen tilaston laatijat saavat apua tilastotieteeltä ja muiltakin tieteenaloilta, yliopistojen tutkimukselta ja opetukselta. Virallinen tilasto hyötyy uudenlaisesta monitoimijaisesta yhteistyöstä. Uuden, kehittyvän teknologian oivaltavasta soveltamisesta on iso apu.

Tilastoseura on kahdessa perättäisessä iltapäiväseminaarissa (12.4.2018 ja 6.9.2018) kiinnittänyt huomiota tilastojen muuttuvaan maailmaan, jota teemaa on käsitelty tilastojen käyttäjien, tuottajien, tilastotieteen ja osaamisen näkökulmista. Ensimmäisessä seminaarissa pohdittiin mistä luotettava tieto tunnistetaan ja miten virallinen tilasto palvelee käyttäjiään ja yhteiskuntaa. Toisessa seminaarissa syvennyttiin tarkastelemaan miten tilastot ja tilastotiede haastavat vaihtoehtoisia totuuksia ja mitä tilastojen lukutaito pitää sisällään. Nuorten tilasto-osaamisesta kuultiin monia kiinnostavia ja vakuuttavia esimerkkejä. Seuraavassa luvussa esittelen keskeisiä havaintoja ensimmäisen seminaarin esityksistä.

Tilastot tuovat selkeyttä ja jatkuvuutta muutoksiin ja murroksiin

Tiedosta tekoihin korosti Jenni Airaksinen pohtiessaan, miten tiedon hyödyntämisestä siirrytään tiedolla johtamiseen, yhteisten tulevaisuuksien etsimiseen. Tiedon pitää auttaa näkemään ilmiöt ja niiden ympäristö, auttaa hahmottamaan kokonaisuuksia, muutoksia ja ilmiöiden välisiä riippuvuuksia. Tähän pää-

semiseksi ei välttämättä aina tarvitse suorittaa uusia aineistohankintoja, vaan olemassa olevien aineistojen analysointi ja työstäminen, kommunikointi tilastoja tutkimusyhteisössä sekä keskustelu tiedon käyttäjien kanssa voisivat olla hyvä toimintatapa. Tänä päivänä kilpaillaan näkyvyydestä ja huomiosta. Visuaalisuus valtaa alaa. Visuaalisuus on keino tehdä myös tieto näkyväksi ja houkuttelevaksi. Tiedon visualisointi lisääntyy kovaa vauhtia. Visualisointi on vahva vaikuttamisen keino, joten pitää varoa liian pitkälle menevää ilmiöiden yksinkertaistamista. Olisi tarvetta tiedon jakamisen ympäristön kehittämiseen.

Ari Jaakola kuvasi esityksessään Helsingin uuden kaupunkistrategian kautta kaupunkitilastoon liittyvien tietotarpeiden kehitystä. Selkeä havainto on, että kaupunkitilaston kysyntä kasvaa. Maailman megatrendit, kuten kaupungistuminen, globalisaatio ja sitä myötä voimistunut kilpailu, digitalisaatio, kansainvälinen muuttoliike, väestön ikääntyminen ja ilmastonmuutokset haastavat kaupunkeja. Vertailevan kaupunkitiedon merkitys korostuu. Mutta tämän rinnalla tarvitaan entistä monipuolisemmin tietoa oman kaupungin ja kaupunkiseudun sisäisestä kehityksestä. Uusia tietotarpeita nousee jatkuvasti esille, joten pitää löytää keinot tuottaa tietoa uusista teemoista ja ilmiöistä sekä kyetä tarkastelemaan tietoa myös hyvin paikallisella tasolla. Toisin sanoen tarvitaan tietoa kaupungin eri alueista, vakiintuneista kaupunginosista aina uusiin suunnittelun tai rakentamisen kohteena oleviin alueisiin. Lisäksi tarvitaan tietoa siitä, miten kaupunkitilaa käytetään, väestökehityksestä, paikkojen ja palvelujen saavutettavuudesta sekä siitä, miten kaupunki ylipäätään toimii. Tarpeet ovat moninaiset ja niihin vastaaminen haastavaa. Perustan tiedon hyödynnettävyydelle muodostavat pienaluetilastot ja paikkatiedot. Tässä korostuvat yhteistyö eri tiedontuottajien kesken ja yhteentoimivuus tilastotiedon ja paikkatiedon välillä Helsingin kaupunkistrategian visio on olla ”Maailman toimivin kaupunki” ja sitä kohti edetessä tarvitaan paljon hyvää ja merkityksellistä tietoa.

Timo Koskimäki tarkasteli tilastoa ymmärryksen lisääjänä esitelmässään ”Official Statistics as a Tool for Making Sense”. Tarkastelun viitekehyksenä hän käytti yhtäältä Brenda Dervinin ”Theory of sense-making” (Dervin, 1998), yhtäältä YK:n ”Fundamental Principles of Official Statistics” (UN Economic and Social Council, 2013). Dervinin lähetymistapa korostaa käyttäjälähtöisyyttä tiedonhankinnassa ja -käytössä. Tilastokeskus noudattaa toiminnassaan YK:n hyväksymiä virallisen tilaston perusperiaatteita, jotka määrittelevät virallisen tilaston tuottajien oikeudet ja velvollisuudet yhteiskunnassa sekä tärkeimmät tilastoja laadittaessa ja julkaistaessa noudatettavat periaatteet, kuten luotettavuus ja tietosuoja.

Maria Valaste vei seminaarin osanottajat aikamatkalle: Minkälainen on tilastotieteilijä 2060-luvulla? Mitä työtä hän tekee ja keiden kanssa? Mihin tiedepohjaan ja kontekstiin ankkuroituu? Virisi hyvää keskustelua ja visiointia. Pohdittiin myös tilastotieteen ja muiden tieteiden yhteistyön ja vuorovaikutuksen tarvetta. Erityisesti tilastotieteen ja datatieteen yhteistyötä toivottiin tiiviimmäksi. Aihepiiriin päätettiin palata.

On olemassa erityinen kansainvälinen maisteriohjelma, joka kouluttaa ja valmentaa virallisen tilaston osaajaksi: EMOS, European Master in Official Statistics (UN Economic and Social Council, 2018). Tätä ohjelmaa esitteli EMOS-ohjelman suorittanut Xiaoyuan Li.

Yhteenvetoa ja pohdintaa

Kaikki tarvitsemme ja haluamme hyvää, luotettavaa ja merkityksellistä tietoa. Tähän on pyrittävä eritoten muuttuvissa olosuhteissa. Korkean laadun saavuttaminen edellyttää tiedon tuottajalta monen alueen osaamista (tietosisältö, metodit, teknologiat, kontekstuaaliset tekijät, kommunikointi) ja monen toimijan yhteistyötä (tilaston tuottajat; tiede, tutkimus ja opetus; tilaston käyttäjät; media; uusia aineistoja tuottavat yritykset; teknologiayritykset ja -alustat; kansainvälinen tilastoyhteisö).

Kun haetaan luotettavaa tietoa, kansallinen ja kansainvälinen virallinen tilasto on keskeisessä asemassa. Virallisella tilastolla on tiukat laatuvaatimukset, se tuotetaan riippumattomasti, se tukeutuu sisällöltään ja menetelmiltään tilastotieteeseen, yhteiskunta- ja ympäristötieteisiin, se soveltaa kansainvälisiä standardeja ja hyödyntää testattua tietoteknologiaa.

Uusien aineistojen (Big Data, IoT, citizens science ym.) hyödyntäminen tilastojen laadinnassa ja muussa tietotuotannossa tai tietopalvelussa edellyttää huomion kiinnittämistä erityisesti yksityisyyden suojaan ja eettisiin kysymyksiin. Tätä on paljon käsitelty ESS:n (European Statistical System) piirissä ja 12.8.2018 julkaistiin kannanotto, joka kantaa nimeä ”Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics)” (European Statistical System Committee (ESSC) meeting, 2018).

Kun tiedon saatavuutta ja avoimuutta ja tilastojen lukutaitoa entisestään edistetään, niin saadaan tieto kaikille ja virallinen tilasto vielä paremmin palvelemaan koko yhteiskuntaa.

Kirjallisuus

[1] Brenda Dervin. “Sense-making theory and practice: an overview of user interests in knowledge seeking and use”. *Journal of knowledge management* 2.2 (1998), s. 36–46.

[2] European Statistical System Committee (ESSC) meeting. *Bucharest memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics)*. https://ec.europa.eu/eurostat/documents/7330775/7339482/The+Bucharest+Memorandum+on+Trusted+Smart+Statistics+FINAL.pdf/59a1a348-a97c-4803-be45-6140af08e4d7. 2018.

[3] UN Economic and Social Council. *What is EMOS?* https://ec.europa.eu/eurostat/cros/content/what-emos_en. 2018.

[4] UN Economic and Social Council. *Fundamental Principles of Official Statistics*. https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf. 2013.

Iltapäiväseminaari — Tilastot muuttuvassa, vaihtoehtoisten totuuksien maailmassa

6. Syyskuuta 2018
Tilastokeskus, Työpajankatu 13, Helsinki

Ohjelma

13:00 Avaus ja johdanto

Tilastoseuran puheenjohtaja Pauliina Ilmonen, Aalto-yliopisto

Tietopalvelujohtaja Hannele Orjala, Tilastokeskus

13:20 **Miten tilastotiede ja tilastot haastavat vaihtoehtoiset totuudet?**

Yliopistonlehtori Kimmo Vehkalahti, Helsingin yliopisto

Tietopalvelujohtaja Hannele Orjala, Tilastokeskus

14:00 Kahvitauko

14:30 **Luotettavaa tilastoakin on osattava lukea; on tunnettava tiedon käytön mahdollisuudet ja rajoitteet. Case: BKT**

Suunnittelija Katri Soinne, Tilastokeskus

15:00 **Miten innostan nuoria tilasto-osaajiksi?**

Tilasto-olympialaisten voittajanuorten esitys

15:30 Loppukeskustelu ja seminaarin päätös

Iltapäiväseminaari järjestettiin yhteistyössä Tilastokeskuksen kanssa.

Tilastotieteestä valmistuneiden työllisyys: totuus ja tilastot

Juha Karvanen
Jyväskylän yliopisto

Tiivistelmä

Opetusministeriön vipunen.fi-palvelun luvut antavat ensisilmäyksellä täysin väärän käsityksen tilastotieteestä valmistuneiden työllisyydestä. Yhdistämällä raportin vuodet 2009–2016 käy kuitenkin ilmi, että noin 5 % tilastotieteestä valmistuneista on ollut työttömänä vuosi valmistumisen jälkeen. Harhaanjohtavien lukujen syyksi paljastuu raporttiin sovellettu tietosuojaus.

Tilastotieteilijöiksi valmistuvien työllisyydestä välittyy yliopistossa työskentelevälle erittäin myönteinen kuva: työnantajat tiedustelevat lähiaikoina valmistuvista ja useilla opiskelijoilla on oman alan työpaikka jo ennen valmistumista. Silmäys julkaistuihin työllistymistilastoihin antaa asiasta kuitenkin erilaisen käsityksen: opetushallinnon tilastopalvelu Vipunen kertoo, että tilastotieteestä vuonna 2016 valmistuneista peräti 12.5 % on työttömänä vuosi valmistumisen jälkeen (Opetushallinto, 2018a) (lyhytlinkki https://urly.fi/12kJ). Kumpi on lähempänä totuutta: tilasto vai tilastotieteilijöiden näppituntuma?

Valmiiksilaskettujen prosenttien sijaan kannattaa tarkastella valmistuneiden lukumääriä, jotka on esitetty Vipusen interaktiivisessa raportissa ”Yliopistosta valmistuneiden työllistyminen” (Opetushallinto, 2018b) (lyhytlinkki https://urly.fi/10Ov). Tilastotieteen luvut tilastovuosilta 2009–2016 on koottu taulukkoon 16.1. Vipusen raporttiselitteen mukaan henkilöt on poimittu aineistoon viimeisimmän ja korkeimman suoritetun tutkinnon mukaan ja raportissa tarkastellaan tilastovuotta edeltävänä vuonna vuosina yliopistotutkinnon suorittaneiden työllistymistilannetta tilastovuoden lopussa. Tietolähteeksi mainitaan opetushallinnon ja Tilastokeskuksen tietopalvelusopimuksen aineisto 4.3.

Taulukossa 16.1 huomio kiinnittyy lähes välittömästi siihen, että luku 3 esiintyy poikkeuksellisen usein ja kaikki taulukon luvut ovat jaollisia kolmella. Vipusen raporttiselitteessä kerrotaan seuraavaa: ”Huom! Raportille on tehty tietosuojaus. Lukumäärätiedot on tasoitettu kolmella jaolliseksi, jolloin pienet, alle viiden lukumäärät näkyvät arvona 3, ja tätä suurempiin arvoihin tulee satunnaisesti lisäys -1, 0 tai +1. Suojauksen tuottamat prosentuaaliset muutokset ovat suurempien arvojen osalta marginaalisia. Esimerkiksi 100:n tapauksen osalta

Taulukko 16.1: Tilastotieteestä valmistuneiden sijoittuminen vuosi valmistumisen jälkeen tilastopalvelu Vipusen mukaan. Luvut on muunnettu kolmella jaollisiksi.

Vuosi	Työllinen	Päätoim. opiskelija	Työtön	Muu	Muuttanut maasta	Yhteensä
2009	36	3	3	0	3	42
2010	9	3	3	0	3	12
2011	24	0	0	0	0	24
2012	36	3	3	0	3	39
2013	30	0	3	3	0	36
2014	24	3	3	0	0	27
2015	24	3	3	3	0	30
2016	18	0	3	3	3	24
2009–2016	201	9	**12**	6	6	**234**

virhe on korkeintaan 1 %, ja 1000:nen tapauksen osalta korkeintaan 0,1 %."

Toisin sanoen symboli "3" sarakkeessa "Työtön" tarkoittaa, että työttömiä on ollut 1, 2, 3, tai 4. Näiden vaihtoehtojen todennäköisyydet eivät ole yhtäsuuret, vaan todellinen työttömien lukumäärä on luultavasti usein ollut 1. Pienten tieteenalojen kannalta pyöristyksen vaikutus on kaikkea muuta kuin marginaalinen. Asialla on merkitystä, koska työllistymistietoja käytetään päätöksenteossa ja niitä vertaillaan julkisuudessa.

Interaktiivisessa raportissa on mahdollista myös valita useita vuosia kerrallaan. Tulosten perusteella vaikuttaa siltä, että tällöin alkuperäiset pyöristämättömät luvut lasketaan ensin yhteen ja pyöristetään sen jälkeen. Valitsemalla vuodet 2009–2016 nähdään, että tilastotieteestä on valmistunut 234 (±1) maisteria, joista 12 (±1) on ollut työttömänä vuosi valmistumisensa jälkeen. Näiden lukujen perusteella tilastotieteestä valmistuneiden työttömyysaste on ollut noin 5 % (tarkempi analyysi sopii vaikkapa harjoitustehtäväksi tilastollisen päättelyn kurssille). Tämä luku on samaa tasoa kuin tekniikan alan ja ICT-alan työttömyys samalla ajanjaksolla ja selvästi pienempi kuin muiden luonnontieteiksi laskettujen koulutusalojen työttömyys.

Tämä esimerkkitapaus osoittaa, että huonosti toteutetut tietosuojakorjaukset voivat vääristää tilastoja kohtalokkaalla tavalla. Tässä tapauksessa "suojaus" ei edes suojaa alkuperäistä tietoa, vaan todellinen työttömien lukumäärä on usein pääteltävissä erilaisia vuosikombinaatioita valitsemalla. Esimerkki kertoo omalla tavallaan siitä, että tilastotieteilijöille riittää töitä tulevaisuudessakin.

Raportista vastaavat henkilöt ovat tietoisia pyöristämiseen liittyvistä ongelmista ja pyysivät ehdotuksia tilanteen korjaamiseksi. Esimerkiksi seuraavia toimenpiteitä voisi harkita:

(i) Lukijaa kehotetaan tutustumaan raporttiselitteeseen taulukon otsikossa ja jossakin muussa näkyvässä paikassa.

(ii) Pyöristetyt lukumäärät esitetään lukuväleinä, esim. 1–4 tai 5–7.

(iii) Prosenttiosuudet lasketaan vain, jos jakaja on riittävän suuri (vähintään 100). Jos jakaja on pienempi kuin 1000, kokonaiset prosentit (esim 6 %) ovat sopiva esitystarkkuus.

(iv) Avataan Tilastokeskuksen kanssa keskustelu siitä, mitä todellisia tai kuviteltuja tietosuojariskejä aineistoon liittyy ja kuinka merkittäviä nämä riskit ovat verrattuna täsmällisestä tiedosta saatuun hyötyyn.

Nämä ehdotukset on toimitettu eteenpäin Opetusministeriöön. On siis mahdollista, että raporttiin on tehty korjauksia tämän artikkelin kirjoittamisen jälkeen.

Kirjallisuus

[1] Opetushallinto. *Opetushallinnon tilastopalvelu*. https://vipunen.fi/fi-fi/_layouts/15/xlviewer.aspx?id=$=$/fi-fi/Raportit/Yliopistostavalmistuneidentyöllistyminen-koulutusala(prosentit).xlsb. Luettu: 2018-09-07. 2018.

[2] Opetushallinto. *Opetushallinnon tilastopalvelu*. https://vipunen.fi/fi-fi/_layouts/15/xlviewer.aspx?id=$=$/fi-fi/Raportit/Yliopistostavalmistuneidentyöllistyminen-koulutusala.xlsb. Luettu: 2018-09-07. 2018.

Regional workshop of European young researchers in statistics

Niko Lietzén, Tommi Mäklin
Lokakuu 29-31 2018, Pariisi (Ranska)

Regional workshop of European young researchers in statistics järjestettiin ensimmäistä kertaa Pariisissa Henri Poincaré -instituutin tiloissa 29 - 31.10.2018. Tapahtuman tavoitteena oli, kuten jo nimestä voi mahdollisesti päätellä, olla tilaisuus, jossa eurooppalaisiksi tilastotieteen harjoittajiksi itsensä kokevat nuoret voivat tutustua toisiinsa. Toisena kunnianhimoisena tavoitteena oli perustaa verkosto laajemmalle eurooppalaiselle yhteistyölle sekä suunnitella tulevia tilastotieteellisiä kuvioita.

Tapahtumaan osallistui yli 20 nuorta tilastotieteilijää kahdeksasta eri maasta paikallisten tilastoseurojen edustajina. Osallistujat olivat pääosin tohtoriopiskelijoita, mutta mukaan mahtui myös muutamia maisteriopiskelijoita sekä vastavalmistuneita tohtoreita ja yksityisellä sektorilla työskenteleviä tieteenharjoittajia. Tohtoriopiskelijat Niko Lietzén Aalto-yliopiston Perustieteiden korkeakoulusta ja Tommi Mäklin Helsingin yliopistosta edustivat itsensä lisäksi tapahtumassa Suomea ja Suomen Tilastoseuraa.

Kolmipäiväinen tutkimustyöpaja koostui tieteellisestä ja hallinnollisesta sisällöstä, joiden tavoitteena oli antaa osallistujille tilaisuus esitellä omia tutkimustuloksiaan vertaisilleen ja toisaalta keskustella siitä, millaisia verkostoja kunkin maan tieteellisillä seuroilla on nuorille tilastotietelijöille. Sekä Niko että Tommi esittivät tapahtuman tieteellisessä osuudessa menestyksekkäästi tutkimustuloksiaan.

Hallinnollisissa istunnoissa päätettiin luoda uusi eurooppalainen verkosto nuorille tilastotieteilijöille sekä tavoitella vastaavan jokavuotisen tapahtuman järjestämistä. Erityisesti ideoitiin sitä, miten kommunikaatiota eri maissa vaikuttavien nuorten tilastotieteilijöiden välillä voitaisiin parantaa ja millä keinoin luotu verkosto tavoittaisi mahdollisimman laajan yleisön. Pähkäilyn tuloksena päätettiin aloittaa verkostoituminen luomalla yhteisölle omat kanavat erinäisiin

sosiaalisiin medioihin sekä järjestää vuonna 2019 kaikille itsensä nuoriksi eurooppalaisiksi tilastotieteilijöiksi kokeville avoin vastaava tapahtuma Romanian pääkaupungissa Bukarestissa.

Kolmipäiväinen tapahtuma oli täynnä virallista ohjelmaa, mutta ystävällisen järjestäjätoimikunnan johdolla aikaa jäi tutustua syksyiseen Pariisiin ranskalaisen ruokakulttuurin muodossa. Tapahtuma oli erittäin lämminhenkinen ja motivoiva ja kaikkia nuorekkaita tilastotieteilijöitä suositellaankin liittymään luotuihin verkostoihin ja harkitsemaan osallistumista seuraavaan tapaamiseen Bukarestissa.

Suomen Tilastoseuran hallitus vuonna 2017

Board members of the Finnish Statistical Society in 2017

Puheenjohtaja Chair	Jyrki Möttönen	Filosofian tohtori Ph.D.
Varapuheenjohtaja Vice Chair	Ari Jaakola	Filosofian maisteri M.Sc.
Rahastonhoitaja Treasurer	Emma Kämäräinen	Valtiotieteiden kandidaatti B.Soc.Sc.
Sihteeri Secretary	Tommi Mäklin	Filosofian maisteri M.Sc.
Jäsen Member	Paula Bergman	Filosofian maisteri M.Sc.
Jäsen Member	Tommi Härkänen	Filosofian tohtori Ph.D.
Jäsen Member	Tara Junes	Valtiotieteiden maisteri M.Soc.Sc.
Jäsen Member	Pihla Oksanen	Valtiotieteiden ylioppilas Student of Soc.Sc
Jäsen Member	Pekka Pere	Doctor of Philosophy D.Phil.
Jäsen Member	Johanna Seppänen	Filosofian tohtori Ph.D.

Suomen Tilastoseuran hallitus vuonna 2018

Board members of the Finnish Statistical Society in 2018

Puheenjohtaja Chair	Pauliina Ilmonen	Filosofian tohtori Ph.D.
Varapuheenjohtaja Vice Chair	Ari Jaakola	Filosofian maisteri M.Sc.
Rahastonhoitaja Treasurer	Mikhael Koufos	Valtiotieteiden maisteri M.Soc.Sc.
Sihteeri Secretary	Tommi Mäklin	Filosofian maisteri M.Sc.
Jäsen Member	Paula Bergman	Filosofian maisteri M.Sc.
Jäsen Member	Tommi Härkänen	Filosofian tohtori Ph.D.
Jäsen Member	Jyrki Möttönen	Filosofian tohtori Ph.D.
Jäsen Member	Pekka Pere	Doctor of Philosophy D.Phil.
Jäsen Member	Johanna Seppänen	Filosofian tohtori Ph.D.
Jäsen Member	Marianne Laalo	Valtiotieteiden kandidaatti B.Soc.Sc.

Gunnar Modeen -minnesmedaljen

Jukka Hoffrén

Statistiska Samfundet i Finland r.f. har i samband med de nordiska statistikdagarna traditionsenligt delat ut Gunnar Modeen -minnesmedaljen till särskilt meriterade sta-tistiker. Praxisen har varit att dela ut medaljen till en representant för det land där statistikdagarna hålls.

Gunnar Modeen -minnesmedaljen beviljas för en betydande livsgärning inom sta-tistikbranschen. Meningen är att den person som belönas är en framstående senior expert inom statistikbranschen, som uttryckligen utmärkt sig i det praktiska statistik-arbetet och som uppskattas av sina kolleger.

Styrelsen för Statistiska Samfundet väljer den person som får medaljen och medaljen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga överlåts medaljen till en betydande nordisk statistiker från det land som respektive år arran-gerar mötet. Den första medaljen överläts vid det nordiska statistikermöte som hölls i Finland år 1989.

Bakgrunden till och kriterier för GM-minnesmedaljen

Efter Gunnar Modeens bortgång år 1988 grundades en medaljfond till hans minne. Medaljen utarbetades på basis av den medaljong som Gunnar Modeens familj gett konstnären Matti Haupt i uppdrag att utforma till Modeens 70-årsdag år 1965. Mot-tagaren av medaljen väljs av styrelsen för Statistiska Samfundet i Finland och medal-jen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga över-låts medaljen till en betydande nordisk statistiker från det land som respektive år ar-rangerar mötet. Den första medaljen överläts vid Nordiska Statistikermötet i Finland år 1989. Priset utdelas vart tredje år till en meriterad statistiker från det land där Nordiska Statistikermötet anordnas.

Allmänna kriterier för Gunnar Modeen -minnesmedaljen:

- priset beviljas för en betydande livsgärning inom statistikbranschen.

Den person som tilldelas medaljen:

- är en expert inom statistikbranschen, som uttryckligen utmärkt sig i det praktiska statistikarbetet
- är en nordisk, framstående senior expert som uppskattas av sina kolleger,

– har akademisk examen (magister, licentiat eller doktor) och

- är villig att ta emot GM-medaljen

Mottagare av GM-minnesmedaljen

Den första medaljen tilldelades Mauno Koivisto, Finlands dåvarande president, som en särskild hedersbetygelse. År 1989 var han beskyddare av Nordiska Statistikermö-tet i Finland som firade 100-årsjubileum för nordisk statistik. Ytterligare en medalj delades ut på mötet och mottagare var professor Eino H. Laurila. Övriga mottagare av medaljen:

År 1992 tilldelades medaljen inte.

År 1995 direktör Poul Jensen, Danmarks Statistik.

År 1998 professor Sven Nordbotten, Universitetet i Bergen.

År 2001 professor Emeritus Gunnar Kulldorf, Umeå universitet.

År 2004 direktör Asta Manninen, Helsingfors stads faktacentral.

År 2007 generaldirektör Hallgrímur Snorrason, Hagstofa, Island.

År 2010 direktör Lars Thygesen, Danmarks Statistik.

År 2013 Liv Hobbelstad Simpson, pensionerad från Statistisk sentralbyrå (SSB) som Head of National accounts och past chair of IARIW

År 2016 Eva Elvers, PhD, pensionerad från Design and Plan & Build and Test som Process owner

Scandinavian journal of statistics

Recognised as a leading journal in its field, the Scandinavian Journal of Statistics is an international publication devoted to reporting significant and innovative original contributions to statistical methodology — both theory and applications. The journal specializes in statistical modelling, showing particular appreciation of the underlying substantive research problems. Scandinavian Journal of Statistics is published on behalf of the Danish Society for Theoretical Statistics, the Finnish Statistical Society, the Norwegian Statistical Society, and the Swedish Statistical Society. The journal is currently edited by professors Peter Dalgaard and Niels Richard Hansen. The national editor for Finland is Jukka Corander (University of Helsinki, Finland), and the other national editors are Jacob von Bornemann Hjelmborg (University of Southern Denmark, Denmark), Geir Olve Storvik (University of Oslo, Norway), and Jimmy Olsson (KTH Royal Institute of Technology, Sweden). The chairman of the board is Thomas Scheike (University of Copenhagen, Denmark) and the board members are Juha Karvanen (University of Jyväskylä, Finland), Hans Karlsen (University of Bergen, Norway), and Sara Sjöstedt de Luna (Umeå University, Sweden).

Scandinavian Journal of Statistics is published quarterly in March, June, September and December by Wiley-Blackwell Publishers, 108, Cowlwy Road, Oxford OX4, 1JF, UK or 238 Main Street, Cambridge, MA 02142, USA.

Members of the Finnish Statistical Society are entitled to discount prices when ordering the Scandinavian Journal of Statistics. For further information, please see the webpage at http://www.wiley.com/bw/subs.asp?ref=0303-6898&site=1

ISI Journal Citation Reports® Ranking: 2018: 64/123 (Statistics & Probability).
Impact Factor: 1.017.
Online ISSN: 1467-9469.

Myönnetyt palkinnot

Leo Törnqvist –palkinnot

1978 Rene Tigerstedt, Helsingin yliopisto. En modell för valbeteende i trafiken.

1979 Pirkko Kirjavainen, Turun kauppakorkeakoulu. Mallin rakentaminen ja ennusteen laatiminen Suomen sähkön kulutukselle kahta aikasarja-analyysimenetelmää käyttäen.

1980 Esa Läärä, Helsingin yliopisto. Ikä-, aika- ja kohorttitekijöiden vaikutukset Suomen miesten keuhkosyöpäsairastavuudessa vuosina 1953–76.

1981 Arvi Suvanto, Tampereen yliopisto. Kausivaihtelu aikasarjamalleissa.

1982 Maija Salo, Helsingin yliopisto. Yritys prioritiedon käytöstä alkoholijuomien kulutusta selittävän kysyntämallin tukena. Jamel Boucelham, Jyväskylän yliopisto: Tunnustuspalkinto.

1983 Vesa Vihriälä, Helsingin yliopisto. Aikasarjojen välisen riippuvuuden mittaus ja testaus: sovellus suomalaisiin rahatalouden sarjoihin. Pirkko Welin, Tampereen yliopisto: Tunnustuspalkinto.

1984 Jari Palsio, Turun kauppakorkeakoulu. Skenaarioiden rakentaminen ristivaikutusanalyysimallia käyttäen.

1985 Kenneth Nordström, Helsingin yliopisto. Gauss-Markov-mallien erikoisongelmista.

1986 Tapio Nummi, Tampereen yliopisto. APL-pohjainen ohjelmisto GMANOVA-mallille.

1987 Ari Veijanen, Helsingin yliopisto. Pickardin kentän soveltamisesta kuvaanalyysissä. Kari Nissinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1988 Jaason Haapakoski, Helsingin yliopisto. Binomijakautuneiden muuttujien muutospisteongelma.

1989 Pasi Korhonen, Helsingin yliopisto. Kemometrian tilastollisista menetelmistä.

1990 Päivi Partanen, Jyväskylän yliopisto. Suljetun populaation koon estimointi merkintä-takaisinpyynti-menetelmällä: log-lineaarinen lähestymistapa. Markku Nurhonen, Tampereen yliopisto: Tunnustuspalkinto.

1991 Elina Järvinen, Helsingin yliopisto. Rajoitettujen, stokastisten ja konveksien estimaattoreiden käytöstä polynomisen viipymämallin parametrien estimoinnisssa simulointikokeiden valossa.

1992 Jouni Kuha, Helsingin yliopisto. Binääristen regressiomallien selittäjien mittausvirheet ja parametriestimaattien mittausvirhekorjaukset. Juha Heikkinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1993 Palkintoa ei jaettu (yhtään ehdotusta ei saatu).

1994 Ilkka Taskinen, Jyväskylän yliopisto. Äärelliset Markovin ketjut ja annelointi.

1995 Mika Rautakorpi, Teknillinen korkeakoulu. Application of Markov chain techniques in certification of software. Tuija Jäppilä, Jyväskylän yliopisto: Tunnustuspalkinto.

1996 Veli-Matti Suppola, Jyväskylän yliopisto. Robustit menetelmät. Jakaumien vinouden vaikutuksesta korrelaatiomatriisin estimointiin.

1997 Albert Höglund, Teknillinen korkeakoulu. An Anomaly Detection System for Computer Networks.

1998 Samuli Visuri, Oulun yliopisto. Robustista kovarianssimatriisin estimoinnista ja sen sovelluksista signaalinkäsittelyssä.

1999 Jani Raitanen, Tampereen yliopisto. Jalkapallo-ottelun lopputuloksen tilastollinen mallintaminen.

2000 Reijo Sund, Helsingin yliopisto. Tilastollisia menetelmiä dynaamisten potilaspopulaatioiden mallintamiseen. Tapahtumahistoria-analyysia hoitoilmoitusrekisterin skitsofreenikoille.

2001 Samu Mäntyniemi, Oulun yliopisto. A Hierarchial Bayes Model for Assessing Salmon (Salmo salar L.) Parr and Smolt Populations.

2002 Ilmari Juutilainen, Oulun yliopisto. Teräslevyjen lujuuden ennustaminen regressio- ja neuroverkkomalleilla.

2003 Leena Kalliovirta, Helsingin yliopisto. Mar-malli.

2004 Mikko Myrskylä, Jyväskylän yliopisto. Estimation of Class Frequencies with Micro Level Auxiliary Information.

2005 Antti Liski, Tampereen yliopisto. Lonkkamurtumapotilaiden hoitokustannusten vertailu vastaavuuspistemäärään perustuvalla menetelmällä.

2006 Karri Seppä, Oulun yliopisto. Suomalaisten paksusuolisyöpäpotilaiden ennusteen analyysi suhteellisen elossapysymisen ja syykohtaisen kuolleisuuden malleilla käyttämällä suurimman uskottavuuden ja Bayesin menetelmiä.

2006 Jukka Siren, Helsingin yliopisto. Populaatioiden geneettisen rakenteen spatiaalinen mallintaminen.

2007 Outi Ahti-Miettinen, Helsingin yliopisto. Kaksivaiheisen potenssikiintiöinnin käyttö otoksen tehostamisessa - Esimerkkinä otoksen suunnittelu työvoimakustannusindeksin tietojen keruulle.

2008 Paul Catani, Svenska handelshögskolan. Enhetsrottest och initialvärdet Tillämpning på arbetslösheten i Finland

2009 Elina Ahola, Jyväskylän yliopisto. Eksponenttisen perheen tila-avaruusmallien sovellus alkoholikuolleisuusaineistoon Matias Leppisaari, Aalto yliopiston teknillinen korkeakoulu: Tunnustuspalkinto.

2010 Sanna Peltomäki, Tampereen yliopisto. Estimation of Below Threshold Intra-EU Trade.

2011–2012 Tytti Pasanen, Tampereen yliopisto. Two-Level Structural Equation Modeling with Non-Normal Observed Variables for Assessing Poverty in Laos.

2013–2014 Joni Virta, Turun yliopisto. Some tools for linear dimension reduction.

2015-2016 Niko Lietzén, Aalto-yliopisto. New Approach to Complex Valued ICA: From FOBI to AMUSE

2015-2016 Santtu Tikka, Jyväskylän yliopisto. Kausaalivaikutusten identifiointi algoritmisesti

Väitöskirjapalkinnot

2009-2012 Jukka Sirén, Helsingin yliopisto. Statistical models for inferring the structure and history of populations from genetic data.

2013-2016 Johan Pensar, Åbo Akademi. Structure Learning of Context-Specific Graphical Models.

Suomen Tilastoseuran julkaisuja

Publikationer utgivna av Statistiska Sammanfundet
Publications issued by the Finnish Statistical Society

Suomen Tilastoseuran julkaisuja

1. Monikielinen väestötieteen sanakirja, suomenkielinen laitos, Helsinki 1962.
   Multilangual Demographic Dictionary, Finnish section, Helsinki 1962.

2. Suomen Tilastoseura – Statistiska Sammanfundet i Finland 1920-1970, Porvoo - Borgå 1970.

3. Pohjoismainen tilastosanasto, toinen tarkistettu laitos.
   Nordisk statistik nomenklatur, andra reviderade upplagan.
   Nordic statistical nomenclature, 2nd revised edition. Jyväskylä 1975

4. Aikasarja-analyysin menetelmiä, Helsinki 1977.

5. Pekka Tavaila: Leo Törnqvist Posti- ja lennätinhallituksen liiketaloudellisen tutkimuslaitoksen esimiehenä 1949–1977, Helsinki 1982.

6. Otanta teoriassa ja käytännössä. Vesa Kuusela ja Leif Nordberg (toim.). Helsinki 1986.

7. Suomen Tilastoseura 70 vuotta. Statistiska Sammanfundet i Finland 70 år.
   The Finnish Statistical Society 70 years. Helsinki 1991.

Tilastotieteellisiä tutkimuksia

Statistiska undersökningar

Statistical Research Reports

ISSN 0356-3499

1. Pentti Manninen: Puolueiden kannatusosuuksien estimoinnin tarkkuus Demingin vyöhykepoiminnassa.
   The Accuracy of Party Support Estimation in Deming Zone Selection.
   (In Finnish with English Summary). Helsinki 1976.
2. Timo Hakulinen: On Competing Risks of Death. Helsinki 1977.
3. Lars-Erik Öller: Time Series Analysis of Finnish Foreign Trade. Helsinki 1978.
4. Pekka Laippala: The Empirical Bayes Two-Action Rules with Floating Optimal Sample Size and Exponential Conditional Distributions. Helsinki 1980.
5. Markku Nurminen: Some Developments in Quantitative Methods of Epidemiology. Helsinki 1982.
6. Pentti Saikkonen: Comparing Asymptotic Properties of Some Tests Used in the Specification of Time Series Models. Helsinki 1985.
7. Lauri Tarkkonen: On Reliability of Composite Scales. Helsinki 1987.
8. Juni Palmgren: Models for Categorical Data with Errors of Observation. Helsinki 1987.
9. Ari Veijanen: On Estimation of Parameters of Partially Observed Random Fields and Mixing Processes. Helsinki 1989.
10. Ritva Luukkonen: On Linearity Testing and Model Estimation in Non-Linear Time Series Analysis. Helsinki 1990.
11. Hely Salomaa: Factor Analysis of Dichotomous Data. Helsinki 1990.
12. Kenneth Nordström: Contributions to the Comparison of Linear Models and to the Löwner-Ordering Antitonicity of Generalized Inverses. Helsinki 1990.
13. Seppo Laaksonen: Handling Household Survey Nonresponce Data. Helsinki 1992.
14. Mervi Eerola: On Predictive Causality in the Statistical Analysis of a Series of Events. Helsinki 1993.
15. Mikael Linden: Studies in Integrated and Co-Integrated Economic Time Series. Helsinki 1995.
16. Tadeusz Dyba: Precision of Cancer Incidence Predictions Based on Poisson Distributed Observations. Helsinki 2000.
17. Kimmo Vehkalahti: Reliability of Measurement Scales. Helsinki 2000.
18. Sirpa Heinävaara: Modelling survival of patients with multiple cancers. Helsinki 2003.

Suomen Tilastoseuran vuosikirja

Årsbok för Statistiska Sammanfundet i Finland

The Yearbook of the Finnish Statistical Society

ISBN 0355-5941

1975 Helsinki 1976

1976 Helsinki 1977

1977 Helsinki 1978

1978 Helsinki 1979

1979 Helsinki 1980

1980 Helsinki 1981

1981 Helsinki 1982

1982 Helsinki 1983

1983 Helsinki 1984

1984 Helsinki 1985

1985 Helsinki 1986

1986 Helsinki 1987

1987 Helsinki 1988

1988–1989 Helsinki 1990

1990 Helsinki 1991

1991 Helsinki 1992

1992 Helsinki 1993

1993 Helsinki 1994

1994 Helsinki 1995

1995 Helsinki 1996

1996 Helsinki 1997

1997 Helsinki 1998

1998 Helsinki 1999

1999–2000 Helsinki 2000

2001 Helsinki 2002

2002 Helsinki 2003

2003 Helsinki 2004

2004 Helsinki 2005

2005 Helsinki 2006

2006 Helsinki 2007

2007 Helsinki 2008

2008 Helsinki 2009

2009 Helsinki 2010

2010 Helsinki 2011

2011–2012 Helsinki 2012

2013–2014 Helsinki 2014

2015-2016 Helsinki 2017

Tilastoseuran julkaisuja voi tiedustella sihteeriltä sähköpostitse osoitteesta suomentilastoseura@gmail.com. Joidenkin julkaisujen painokset ovat tosin jo loppuneet.

Muita julkaisuja

Andra publikationer

Other publications

Suomen tilastoseura 1920–1945, Helsinki 1946
Statistiska Sammanfundet i Finland 1920–1945, Helsingfors 1946

Pohjoismainen tilastosanasto – Nordisk statistisk nomenklatur, Kööpenhamina 1954

13:e Nordiska statistikermötet i Helsingfors 14–16 juni 1973, Jyväskylä 1974
The 13th Joint Meeting of the Nordic Statistical Societies in Helsinki June 1973, Jyväskylä 1974

Det 18:e nordiska statistikmötet i Esbo, Hundraårsjubileum, Helsingfors 1990

The Joint Conference of the Nordic Statisticians in Espoo, Finland 1989, Helsinki 1990

ISSN 0355 – 5941