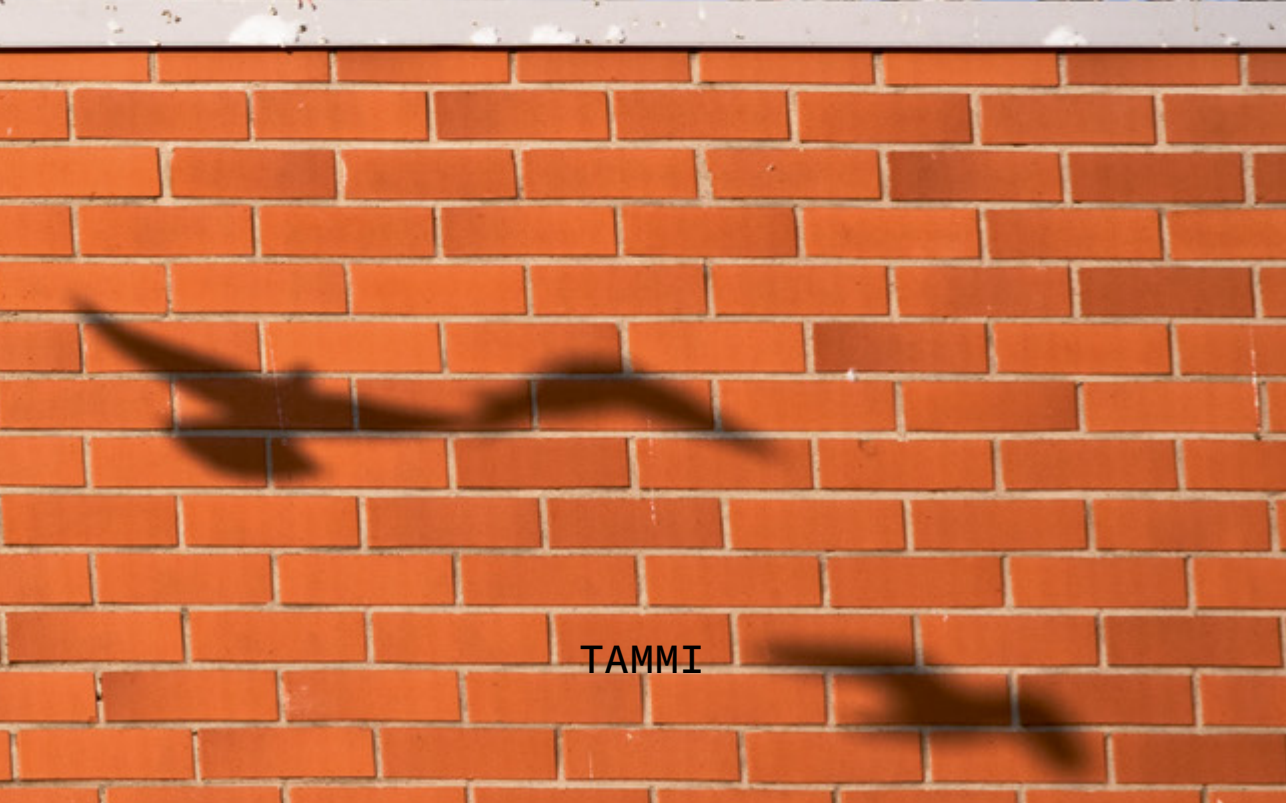




LAURI NUMMENMAA

TILASTOTIETEEN KÄSIKIRJA



TAMMI

TILASTOTIETEEN
KÄSIKIRJA

TAMMI

© Lauri Nummenmaa ja Tammi, 2021
Tammi on osa Werner Söderström Osakeyhtiötä
Graafinen suunnittelu ja taitto Atte Kalke, Vitale Ay
ISBN 978-952-04-0138-2
Painettu EU:ssa

SISÄLLYS

Lukijalle	12	3. Mittaaminen ja aineisto	51
1. Epävarmuuden vähentäminen	15	Mittaaminen, havainnot ja muuttujat.....	51
Järjestystä kaaoksen keskellä.....	15	Havaintomatriisi.....	52
Kuinka näkymätön saadaan näkyväksi?.....	17	Kvalitatiiviset ja kvantitatiiviset muuttujat.....	54
Tutkimustiedon tilastollinen käsittely.....	18	Diskreetit ja jatkuvat muuttujat.....	54
Aineiston kuvaileminen ja yksinkertaistaminen.....	20	Mitta-asteikot.....	55
Muuttujien välisten yhteyksien tutkiminen.....	23	Laatueroasteikko eli nominaaliasteikko.....	55
Päätely ja yleistäminen.....	25	Järjestysasteikko eli ordinaaliasteikko.....	56
Mallintaminen ja ennusteiden laatiminen.....	26	Välimatka-asteikko eli intervalliasteikko.....	57
Ryhmittely ja luokittelu.....	27	Suhdeasteikko.....	57
Laskennan vallankumous.....	29	Absoluuttinen asteikko.....	57
Miksi laskea käsin?.....	29	Mitta-asteikon valinta.....	58
Tilastolliset ohjelmistot.....	30	Mitta-asteikoiden invarianssi.....	59
2. Tieteellisen tutkimuksen perusteet	31	Muiden mitta-asteikoiden invarianssi.....	60
Tieteellinen ajatteleminen.....	31	Miksi invarianssi on tärkeää?.....	61
Empiirinen ja teoreettinen tutkimus.....	34	Populaatio ja otos.....	61
Tieteellisen tutkimuksen peruserä ja oivalluksen kehä.....	36	Otoksen edustavuus ja otantasuhde.....	62
Havaintoaineiston dokumentoiminen.....	38	Yksinkertainen satunnaisotanta.....	63
Havaintoaineiston säilyttäminen.....	39	Systemaattinen otanta.....	63
Tutkimuksen toistettavuus.....	39	Ositettu otanta.....	64
Tutkimusasetelmat.....	41	Ryväsotanta eli klusteriotanta.....	65
Riippumattomat ja riippuvat muuttujat.....	41	Harkinnanvarainen otanta näytteen poimimiseksi.....	65
Ulkoisten tekijöiden kontrolloiminen.....	43	Otoksen koko.....	66
Erilaiset tutkimusasetelmat.....	44	4. R-ohjelman peruskäyttö	69
Kokeellinen tutkimus.....	44	Millainen ohjelma R on?.....	69
Kvasikokeellinen tutkimus.....	45	Tilastollinen ohjelmointi.....	71
Korrelatiivinen tutkimus.....	47	R-ohjelman peruskäyttö.....	72
Pitkittäistutkimus.....	47	Ennen kuin aloitat: hakemiston asettaminen.....	73
Retrospektiivinen tutkimus ja avoimet tietovarannot.....	48	Hyödyllisiä peruskomentoja.....	73
Sopivan tutkimusasetelman valinta.....	48	Peruslaskutoimitukset ja muuttujien käyttäminen.....	74
		Tulostusten tallentaminen.....	76
		Vektorien ja matriisien käsittely.....	76
		Funktioiden käyttäminen.....	80

Tietokehikon luominen aineiston analysoimista varten.....	80
Pakettien lataaminen.....	83
Istunnon hallinta ja skriptien kirjoittaminen	83
Opettele kirjoittamaan selkeää koodia	85
Apua!.....	85

5. Tutkimusaineiston peruskäsittely R:llä 87

Tilastollisen analyysiprojektin järjestäminen	87
Vektorityypit	88
Aineiston syöttäminen R-konsolin avulla.....	89
Kirjan esimerkkiaineistojen lataaminen.....	90
Aineiston lataaminen RStudio valikoiden avulla	91
RStudio dataikkuna	92
Aineistojen lataaminen R-komentojen avulla	93
Excel-tiedostojen lataaminen.....	96
Aineistojen tallentaminen	97
Aineiston liittäminen.....	97
Tietokehikon muuttujien poistaminen ja lisääminen	98
Muuttujatyyppien käsittely ja muuttaminen	100
Puuttuvien havaintojen käsittely.....	101

6. Aineiston kuvaileminen numeerisesti 103

Tutkimusaineiston tiivistäminen taulukoiiksi.....	103
Tilastolliset tunnusluvut: Signaalin paikallistaminen kohinasta	104
Parametrit ja estimaatit.....	107
Frekvenssit	108
Sijaintiluvut.....	108
Moodi (mo).....	108
Mediaani (md).....	108
Alakvartiili (Q1) ja yläkvartiili (Q3).....	109
Fraktiilit.....	110
Aritmeettinen keskiarvo (\bar{x} , M)	110
Geometrinen keskiarvo	112
Harmoninen keskiarvo	114
Hajontaluvut	114
Kvartiiliväli ja vaihteluväli.....	115
Varianssi (s^2 , σ^2)	116
Keskihajonta (σ , s , SD)	117
Variaatiokerroin (v).....	118
Vinous (g_1).....	119
Huipukkuus (g_2)	119
Sijainti- ja hajontalukujen tulkitseminen	120
Kuvailevien tunnuslukujen laskeminen R:llä	121
Sijaintilukujen laskeminen	122
Hajontalukujen laskeminen	123
Tunnusluvut useille muuttujille	124

7. Tilastografiikka 127

Kuva kertoo enemmän kuin tuhat numeroa	127
Maailman kuuluisin kuvaaja.....	128
Sopivan tiivistä grafiikkaa	129
Yleisimmät kuvaajatyytit	130
Pylväskuvaajat	131
Palkkikuvaaja.....	132
Pylväskuvaaja jatkuville muuttujille	133
Viivadiagrammi	134
Viivadiagrammi ja pistekuvaaja jatkuville muuttujille	135
Aluekuvaajat	136
Sektorikuvaaja.....	137
Palkkikuvaajan käyttö frekvenssijakauman kuvaamiseen.....	138
Histogrammi	138
Laatikko-jana-kuvaaja.....	139
Hyvän tilastografiikan perusteet	141
Myös alkuperäisiä havaintoja pitää tarkastella kuvaajien avulla	142
Kuvaajien piirtäminen on vaiheittainen ja kokeilua vaativa prosessi	144
Tehokas kuvaajien käyttäminen.....	145

8. Perusografiikka R-ohjelmistolla 147

Grafiikan laatiminen R-ohjelmistolla.....	147
Ensimmäiset kuvaajat plot -funktion avulla	148
Kuvaajien katseleminen ja tallentaminen	151
Kuvien piirtämisen perusfunktiot	152
Histogrammit.....	153
Laatikko-jana -kuviot.....	154
Pylväskuviot ja sektorikuviot.....	155
Hankalaa grafiikkaa	155
Ggplot2 grafiikan perustyökaluna	156
Ensimmäiset kuvaajat qplotin avulla	156
Pylväskuvaajat qplot-funktiolla	158
Histogrammit ja estimoidut tiheysfunktiot qplot-funktiolla	159
Laatikko-jana-kuvaajat qplot-funktiolla	161
Ggplot2:n käyttäminen	164
Histogrammit ggplot-funktiolla	164
Laatikko-jana -kuviot ggplot-funktiolla	166
Jatkuvien muuttujien tunnuslukujen kuvaaminen ggplot2:n avulla	168
Sirontakuvio ggplot2:n avulla	169
Kategoristen muuttujien kuvaaminen ggplot2:n avulla.....	171

9. Todennäköisyyslaskenta 173

Epävarmuuden hallitseminen todennäköisyyksien avulla	173
Mitä on todennäköisyys?.....	174
Todennäköisyyslaskennan perusaksioomat	175
Subjekttiivinen todennäköisyys	175

Klassinen todennäköisyys	176
Empiirinen lähestymistapa	178
Tilastollinen lähestymistapa	178
Suurten lukujen laki.....	179
Empiirinen tutkimus toistokokeena.....	180
Joukko-opin perusteet.....	181
Kombinatoriikkaa	183
Tuloperiaate	184
N-kertoma	185
Variaatiot ja kombinaatiot.....	185
Todennäköisyyslaskennan laskusääntöjä	186
Erillisten tapahtumien yhteenlaskusääntö	186
Yleinen yhteenlaskusääntö.....	187
Komplementtitapahtuman todennäköisyys.....	188
Riippumattomien tapausten kertolaskusääntö ..	188
Yleinen kertolaskusääntö	189

10. Tilastolliset todennäköisyysjakaumat 191

Empiiriset jakaumat ja tilastolliset todennäköisyysjakaumat	191
Satunnaismuuttujat.....	193
Todennäköisyysjakaumien kuvaaminen.....	194
Diskreetin jakauman tiheysfunktio	195
Diskreetin jakauman kertymäfunktio	195
Diskreetin satunnaismuuttujan odotusarvo	197
Diskreetin jakauman varianssi ja keskihajonta.....	198
Yleisimpiä diskreettejä jakaumia	198
Binomijakauma	198
Binomijakauman todennäköisyysfunktio	200
Binomijakauman kertymäfunktio.....	200
Binomijakauman odotusarvo ja keskihajonta	201
Binomijakaumaan liittyvät R-funktiot	202
Poisson-jakauman todennäköisyysfunktio.....	204
Binomijakauman approksimointi Poisson-jakaumalla	205
Poisson-jakaumaan liittyvät R-funktiot	206
Jatkuvat satunnaismuuttujat	208
Jatkuvan satunnaismuuttujan tiheysfunktio.....	208
Jatkuvan satunnaismuuttujan kertymäfunktio	209
Jatkuvan satunnaismuuttujan odotusarvo ja hajonta.....	210
Normaalijakauma.....	210
Standardoitu normaalijakauma	211
Normaalijakaumaan liittyvät R-funktiot.....	213
Studentin t-jakauma	215
Studentin t-jakaumaan liittyvät R-funktiot	216
χ^2 -jakauma.....	217
χ^2 -jakaumaan liittyvät R-funktiot.....	218
F-jakauma.....	220
F-jakaumaan liittyvät R-funktiot.....	221

11. Tilastollinen mallintaminen ja epävarmuuden vähentäminen 223

Tilastolliset mallit ennustamisen apuna.....	223
Keskiarvo: Yksinkertaisin tilastollinen malli	227
Otantavirhe	228
Keskeinen raja-arvolause.....	230
Miksi keskeinen raja-arvolause on tärkeä?.....	233
Luottamusvälit estimoinnin työkaluna	233
Keskiarvon luottamusvälin estimoiminen normaalijakauman avulla	234
Keskiarvon luottamusvälin estimoiminen t-jakauman avulla.....	236
Luottamusvälin käyttäminen epävarmuuden arvioimisessa	237

12. Tilastolliset testit ja hypoteesien testaaminen 239

Hypoteesien testaaminen.....	239
Tieteelliset hypoteesit.....	242
Tilastolliset hypoteesit	243
Hypoteesien testaaminen tilastotieteen perustyökaluna.....	243
Tilastollinen testi hypoteesien testaamisen työkaluna	245
Yksinkertainen permutaatiotesti R:llä	248
Miksi permutaatiotestauksen lisäksi tarvitaan jakaumamalleja?.....	249
Yleisimmät tilastolliset testityypit.....	250
Havaintojen suuruusluokkaa vertailevat soveltuvat testit	251
Muuttujien välisiä yhteyksiä ja yhteisvaihtelua arvioivat testit	252
Tilastollisten testitulosten arvioiminen p-arvojen avulla.....	252
Keskiarvon luottamusvälit: yksinkertaisin parametrinen tapa hypoteesien testaamiseen ..	254
Yksinkertaisin tilastollinen testi: otoskeskiarvon vertaaminen odotusarvoon Z-testillä	255
Nollahypoteesin testaaminen ja p-arvojen määrittäminen	257
Kahden keskiarvon vertaileminen riippumattomien otosten Z-testillä	260
Virhetyypit	262
Parametriset ja epäparametriset testit	265

13. Aineiston valmisteleminen ja käsittely 267

Digitaaliset tutkimusaineistot	267
Laadunvarmistus	268
Puuttuvat tiedot.....	269
Pudotus.....	269
Korvaaminen	269
Satunnainen vai systemaattinen kato?.....	270

Jakaumien muotojen tarkasteleminen ja muunnosten tekeminen	270
Summamuuttujat ja yhdistelmämuuttujat.....	272
Transformaatioiden tekeminen R:llä	273
Poikkeavat havainnot.....	275
Normaalijakaumaoletuksen testaaminen.....	275
Normaalijakaumaoletuksen testaaminen R:llä	276
Yhteenvedo aineiston valmistelusta.....	279

14. Keskiarvojen vertaileminen t-testeillä 281

Havaintojen suuruusluokan vertaileminen	281
Studentin t-jakauma mallina otoskeskiarvojen jakaumalle	282
Erilaiset t-testit	284
Otoskeskiarvon testaaminen yhden otoksen t-testillä	284
Yhden otoksen t-testin laskeminen	286
Yhden otoksen t-testi R:llä	287
Kahden keskiarvon vertaileminen riippumattomien otosten t-testillä	288
Riippumattomien otosten tutkimusasetelma	290
Kahden keskiarvon vertaileminen riippumattomien otosten t-testillä	291
Riippumattomien otosten t-testin laskeminen	292
Riippumattomien otosten t-testi R:llä	293
Toistomittausasetelmat	297
Toistettujen mittausten t-testi.....	299
Toistettujen mittausten t-testin laskeminen	300
Toistettujen mittausten t-testi R:llä	301
T-testien käyttäminen tutkimusaineiston analysoimisessa.....	303

15. Yksisuuntainen varianssianalyysi 305

Keskiarvojen vertaileminen varianssin avulla.....	305
Mistä mittaustuloksiin aiheutuu vaihtelua?	306
Varianssijajotelmät	308
Fisherin F-suhde	310
F-jakauma	311
Yksisuuntaisen varianssianalyysin laskukaavat.....	313
Varianssianalyysin hypoteesit	313
Neliösummat	313
Neliösummien laskeminen	314
Vapausasteet (df) varianssianalyysissa.....	315
Varianssitermien ja F-suhteen laskeminen	315
Yksisuuntaisen varianssianalyysin laskeminen	316
Vaihe 1: lasketaan tarvittavat keskiarvot	317
Vaihe 2: lasketaan neliösummat	317
Vaihe 3: lasketaan neliösummien vapausasteet	318
Vaihe 4: lasketaan F-suhde	319

Yksisuuntainen varianssianalyysi R:llä.....	319
Varianssianalyysimallin monivertailut	322
Monivertailuissa käytettävät testit.....	325
Kontrastivertailut R:llä	325

16. Useampisuuntainen varianssianalyysi 327

Tarkempia malleja keskiarvojen vaihtelun selittämiseen	327
Päävaikutus ja yhdysvaikutus.....	329
Varianssin hajottaminen useampisuuntaisessa varianssianalyysissa	331
Kaksisuuntainen varianssianalyysimalli	331
Neliösummat kaksisuuntaisessa varianssianalyysissa	332
Vapausasteet kaksisuuntaisessa varianssianalyysissa	333
Varianssitermien ja F-suhteiden laskeminen	334
Kaksisuuntaisen varianssianalyysin laskeminen	335
Vaihe 1: lasketaan tarvittavat keskiarvot	337
Vaihe 2: lasketaan ensimmäiset neliösummat	337
Vaihe 3: lasketaan vapausasteet	339
Vaihe 4: lasketaan varianssit ja F-suhteet	339
Kontrastit ja post hoc -vertailut useampisuuntaisessa varianssianalyysissa.....	340
Kaksisuuntainen varianssianalyysi R:llä	341

17. Toistettujen mittausten varianssianalyysi 345

Toistomittausasetelma	345
Toistettujen mittausten varianssianalyysi	346
Miksi toistomittausmalli on niin tehokas?.....	347
Toistettujen mittausten varianssianalyysin laskukaavat.....	349
Neliösummat	349
Vapausasteet.....	351
Varianssitermien ja F-suhteen laskeminen	351
Toistettujen mittausten varianssianalyysin laskeminen	352
Vaihe 1: lasketaan neliösummat.....	353
Vaihe 2: lasketaan vapausasteet	354
Vaihe 3: lasketaan varianssit ja F-suhde.....	354
Kontrastit ja post hoc -vertailut toistettujen mittausten varianssianalyysissa	355
Toistettujen mittausten varianssianalyysi R:llä	355
Useampisuuntainen toistettujen mittausten varianssianalyysi	358
Toistettujen mittausten varianssianalyysi lohkokotekijällä R:llä	360
Kaksisuuntainen toistettujen mittausten varianssianalyysi R:llä.....	361

18. Epäparametriset menetelmät jakaumien sijainnin vertailemiseen

365

Epäparametriset testit	365
Miksi normaalijakaumaa ei voida aina käyttää mallina jakaumien sijainnista?.....	366
T-testien epäparametriset vastineet	367
Mann–Whitneyn U-testi	368
U-testin laskeminen.....	369
U-testi R:llä.....	370
Wilcoxonin merkittyjen järjestykselukujen testi ..	371
Wilcoxonin testin laskeminen	371
Wilcoxonin testi R:llä.....	372
Kruskal–Wallis-testi.....	374
Kruskal–Wallis-testin laskeminen	374
Friedmanin testi.....	377
Friedmanin testin laskeminen	377
Friedmanin testi R:llä	378
Milloin epäparametrisia testejä pitää käyttää?...	380

19. Korrelaatiokertoimet

383

Yhteisvaihtelun tarkasteleminen	383
Kahden jatkuvan muuttujan välisen yhteyden tarkasteleminen	385
Kovarianssi s_{xy}	386
Tulomomenttikorrelaatiokerroin r	387
Tulomomenttikorrelaatiokertoimen laskeminen	389
Selitysaste.....	390
Korrelaatiokertoimen tilastollinen merkitsevyys	390
Korrelaatiokertoimeen liittyvät kriittiset arvot.....	391
Järjestykskorrelaatiokerroin r_s	392
Järjestykskorrelaatiokertoimen laskeminen.....	393
Muita korrelaatiokertoimia	394
Multippelikorrelaatio $R_{z,xy}$	395
Osittaiskorrelaatio	396
Piste-biseriaalinen korrelaatiokerroin r_{pb}	397
Biseriaalinen korelaatiokerroin r_b	398
Tetrakorinen korrelaatiokerroin r_{tc}	399
Korrelaatiomatriisit	400
Korrelaatiokertoimien tulkitseminen	402
Korrelaatiokertoimen tilastollisen merkitsevyyden tulkitseminen	403
Poikkeavat havainnot.....	404
Vähäinen varianssi.....	404
Epälineaariset yhteydet	405
Piilossa olevat yhteydet	407
Kausaalisuhteiden päättelyn ongelma	407
Korrelaatiokertoimet R:llä.....	409
Korrelaatiomatriisit ja sirontakuviot R:llä	411
Osittaiskorrelaatiokerroin R:llä	412
Multippelikorrelaatiokerroin R:llä.....	413

20. Kategoristen muuttujien väliset yhteydet

415

Kategoristen muuttujien väliset yhteydet.....	415
Pearsonin χ^2 -yhteensopivuustesti	416
Pearsonin χ^2 -yhteensopivuustestin laskeminen...	418
Pearsonin χ^2 -yhteensopivuustesti R:llä	418
χ^2 -riippumattomuustesti.....	420
χ^2 -riippumattomuustestin laskeminen	421
χ^2 -riippumattomuustesti R:llä	422

21. Yhtäsuuruustestejä erikoistapauksiin

425

Erilaisia yhtäsuuruuksia	425
Luottamusväli suhteelliselle otokselle	426
Suhteellisen osuuden testaaminen	428
Kahden suhteellisen osuuden testi	430
Varianssien yhtäsuuruuden testaaminen	431
Korrelaatiokertoimien yhtäsuuruuden testaaminen riippumattomilla otoksilla	435

22. Lineaarinen regressioanalyysi

439

Aineiston kuvaaminen lineaarisen mallin avulla.....	439
Regressiosuoran määrittäminen neliösumman minimoinnin avulla	439
Regressiosuoran määrittäminen	442
Useamman selittäjän regressiomallit.....	445
Millaisiin aineistoihin regressioanalyysi soveltuu?.....	445
Regressiomallin selittäjien valinta	446
Mallin tarkastelu.....	449
Selittäjien tarkasteleminen	450
Jäännöstermien tarkasteleminen	452
Dummy-koodaus.....	453
Lineaarinen regressiomalli R:llä.....	454

23. Aikasarja-analyysi

461

Tulevaisuuden ennustaminen menneisyyden perusteella	461
Aikasarja aineistona.....	462
Aikasarjan kuvaaminen graafisesti	463
Aikasarja-analyysin perusidea.....	467
Kuinka komponentit voidaan tunnistaa aikasarjasta?	468
Virhetermin ja kausivaihtelun suodattaminen liukuvan keskiarvon avulla	468
Autokorrelaatio	469
ARIMA-mallit ja aikasarjaennusteiden laatiminen	470
ARIMA-mallin arvioiminen.....	475
ARIMA-mallin muodostaminen R:llä	475
Dekomponointi	476
ARIMA-mallin sovittaminen	477

24. Elinaika-analyysi 483

Elinaika tutkimuskohteena	483
Elossaolofunktio, välttöfunktio ja vaarafunktio	485
Elinaikataulu ja seurantatiedon järjestäminen	486
Kaplan–Meier-menetelmä	488
Kaplan–Meier-analyysi R:llä	488
Logrank-testi	491
Logrank-testi R:llä	493
Coxin regressioanalyysi	495
Coxin regressioanalyysi R:llä	497

25. Mittauksen reliabiliteetti ja validiteetti 499

Miksi mittauksen tarkasteleminen on välttämätöntä?	499
Mittausteoria	500
Epäsuoran mittaamisen periaate	500
Reliabiliteetti ja validiteetti	501
Klassinen testiteoria	502
Klassisen testiteorian aksioomat	505
Mittausvirheen arvioiminen klassisen testiteorian avulla	506
Reliabiliteettikerroin	507
Reliabiliteetin estimointi	508
Kahden rinnakkaisen mittariversion käyttö	509
Test–retest-menetelmä	509
Split–half-menetelmä	510
Sisäisen konsistenssin menetelmä ja Cronbachin α	510
Spearman-Brownin kaava	512
Arvioijien välinen reliabiliteetti	513
Validiteetti	514
Mittauksen validiteetti	515
Kriteerivaliditeetti	517
Mittauksen validiteettikertoimen estimointi	518
Reliabiliteetin ja validiteetin välinen suhde	520
Reliabiliteettianalyysi R:llä	521

26. Reliabiliteetin vaikutus mittaustuloksiin 525

Mittauksen keskivirhe	525
Mittauksen keskivirhe ja reliabiliteetti	526
Reliabiliteetti ja mittaustulosten käyttäminen tilastollisessa päätelyssä	526
Korrelaation attenuaatio	528
Erottusuureiden reliabiliteetti	530
Summamuuttujien reliabiliteetti	531
Palautuminen kohti keskiarvoa	532
Kuinka reliabeli mittarin pitäisi olla?	535

27. Efektikoko 537

Nollahypoteesin merkitsevyytestaus ja efektikoko	537
Efektikoon estimaatit	539
Standardoituun keskiarvojen erotukseen perustuvat efektikoon estimaatit	540
Korrelaatiokertoimeen perustuvat efektikoon estimaatit	544
Efektikoon η -estimaatti	548
Efektikoon estimaattien käyttäminen	550

28. Meta-analyysi 553

Miksi kaikki tutkimustulokset eivät toistu täsmälleen samanlaisina?	553
Tutkimustulosten yhdisteleminen	554
Meta-analyttinen tutkimus	556
Meta-analyysin suunnitteleminen	559
Tutkimusaineiston etsiminen	560
Kirjallisuuden koodaaminen ja efektikokojen laskeminen	561
Meta-analyysin tulosten laskeminen vakiotekijämallin avulla	564
Efektikokojen painottaminen vakiotekijämallissa	564
Meta-analyysin tulosten laskeminen satunnaistekijämallin avulla	566
Painokertoimien laskeminen varianssihajotelmien avulla	568
Satunnaistekijämalli ja metaregressio	570
Meta-analyysin tuloksien esittäminen graafisesti	571
Suppilokuvaajat	572
Yleisiä ongelmia meta-analyysissa	573
Meta-analyysi R:llä	574
Metaregressioanalyysi	578

29. Tilastollinen luokittelu 581

Havaintojen luokittelu	581
Luokkien kuvaaminen piirteiden avulla	582
Logistiset regressiomallit	583
Logistisen regressiomallin oletukset	584
Logistisen regression perusidea	585
Logistisen regressiomallin muodostaminen ja Maximum Likelihood -estimointi	587
Logistisen regressiomallin arvioiminen	589
Luokitteluratkaisun ennustekyvyn arvioiminen	592
Logistinen regressioanalyysi R:llä	594
Mallin arvioiminen	595
Tilastollinen luokittelu logistisella regressioanalyysilla	596
Erotteluanalyysi	598
Erotteluanalyysin peruseriaate	600
Erotteluanalyysin rajoituksia	601
Erotteluanalyysi R:llä	602

30. Ryhmittelyanalyysi 607

Ryhmittelyanalyysi ja ohjaamaton luokittelu.....	607
Millaisia luokkarakenteita	
ryhmittelyanalyysilla voidaan selvittää?	608
Erlaisia tapoja ryhmittelyn suorittamiseksi.....	609
K-keskiarvoryhmittely.....	610
Keskiarvon siirtoryhmittely.....	612
Hierarkkinen ryhmittely.....	613
Ryhmittelyanalyysin ongelmia.....	614
Ryhmittelyanalyysi R:llä.....	615

31. Moniulotteinen asteikointi 619

Ulottuvuuksien tiivistäminen	619
Etäisyysmatriisit	621
Etäisyyksien muodostaminen arviointien perusteella.....	621
Samankaltaisuuksien arvioiminen järjestystehtävien avulla.....	623
Etäisyyksien muodostaminen muuttujien perusteella.....	623
Moniulotteisen asteikointiratkaisun muodostaminen	624
Kuinka monta ulottuvuutta tarvitaan?	625
Moniulotteisen asteikoinnin ongelmia.....	626
Moniulotteisen asteikointi R:llä.....	627

32. Pääkomponenttianalyysi 633

Voiko dataa olla liikaa?	633
Pääkomponenttianalyysi	634
Miten ulottuvuuksien vähentämistekniikat suhtautuvat toisiinsa?	636
Pääkomponenttianalyysin oletukset	637
Pääkomponenttien muodostaminen	637
Pääkomponenttianalyysin tulosten tulkinta	639
Aineiston tiivistäminen pääkomponenttien avulla ennen muita tilastoanalyseja.....	641
Pääkomponenttianalyysi R:llä.....	642
Kasvonpiirteiden tiivistäminen pääkomponenttianalyysin avulla	644

33. Faktoriantalyysi 649

Korrelaatorakenteen analysoiminen	649
Yhteisvaihtelun tiivistäminen faktoreiksi	650
Kuinka faktoriantalyysi toimii?.....	651
Faktoriantalyysin teoria.....	653
Faktorimallin esitys	655
Faktorimallin suunnitteleminen.....	657
Ekstraktointimenetelmät	659
Rotaatio	660
Mallin tulkinta	661
Faktoriantalyysi R:llä	661

34. Rakenneyhtälömallit 665

Aineiston kuvaaminen mallien avulla.....	665
Yhden muuttujan regressiomallit	666
Rakenneyhtälömallit, kausaalimallit ja polkumallit	667
Rakenneyhtälömallien yleiset ominaisuudet	667
Erlaisia rakenneyhtälömalleja	668
Polkumalli havaituille muuttujille.....	668
Mittausmalli.....	670
Rakennemalli.....	670
Mittaus- ja rakennemalli	671
Rakenneyhtälömallien muodostaminen.....	671
Rakenneyhtälömallien sovelluskohteet	673
Rakenneyhtälömallit R:llä	673
Havaittujen muuttujien regressiomalli.....	674
Havaittujen muuttujien polkumalli mediaatiovaikutuksella.....	675
Konfirmatorinen faktorimalli	676
Rakenneyhtälömalli	676
Rakenneyhtälömallin sovittaminen.....	677
Rakenneyhtälömallin arvioiminen	679
Mallin sopivuuden arvioiminen	681
Estimaatien arvioiminen ja jäännöskorrelaatioiden tarkasteleminen.....	682
Modifikaatioindeksien käyttäminen	683

Lähdeluettelo 685**Liitteet 687**

Tutkimusaineistojen kuvaukset.....	689
Aineistojen käyttäminen	689
1. Aggressio	691
2. Älykkyyks.....	693
3. CIRI	694
4. Dopamiini	695
5. Elinaika.....	696
6. Hevi.....	697
7. Hitfeelshit1	698
8. Hitfeelshit2	699
9. Insuliini.....	700
10. Kasvot.....	701
11. Kipu	703
12. MRI	705
13. Persoonallisuus	706
14. Rakenne.....	707
15. Suhteet	708
16. Sääteily	709
17. Syntyvyys	711
18. Tuntemukset	712
19. Työttömyys.....	713
20. Vaalit.....	714
21. Vilja.....	717

Tilastollisia taulukoita 718

LUKIJALLE

Maailma ei koostu numeroista. Luonnon, tekniikan, talouden ja yhteiskunnan ilmiöiden käsitteleminen numeroina kuitenkin helpottaa niiden jäsentämistä. Numeroiden avulla voidaan kuvailla asioiden välisiä suuruuksia, suhteita, yleisyyttä tai todennäköisyyttä. Tilastotieteen tehtävänä on auttaa tällaisten numeeristen päätelmien tekemisessä. Tilastotiedettä tarvitaan kaikilla tieteenaloilla tutkimusaineiston keräämiseen, kuvailemiseen ja mallintamiseen. Lisäksi monet sovellusalat kuten kaupankäynti, media-ala sekä terveydenhoito ja teknologia tarvitsevat tilastotieteen menetelmiä riskien arvioimiseen ja ennusteiden laatimiseen. Tilastotieteen perusteiden hallinta kuuluukin nykyään kaikkien tieteenalojen yleisivistykseen.

Tämä kirja on kirjoitettu tilastotieteen käyttäjille – sekä opiskelijoille että tutkijoille – yhdistäen tilastotieteen teoriaa ja sovelluksia. Kirja on lähtökohtaisesti soveltavan tilastotieteen perusteos, mutta keittokirjamaisen esityksen sijaan asiat esitellään havainnollisesti tilastotieteen teoriaan perustuen. Satunnaisuuden ja todennäköisyyslaskennan perusteista sekä yksinkertaisimmista perusmenetelmistä on laadittu käsinlaskuesimerkit, koska tällaisia menetelmiä on vielä mielekästä harjoitella ilman tietokoneen apua. Monimutkaisemmista menetelmistä on esitetty niiden teoreettinen perusta intuitiivisesti, ja keskitytty analyysiin liittyviin käytännön kysymyksiin sekä tilastollisen ohjelmoinnin perusteisiin. Kirja etenee tutkimuksen tekemisen perusteista tutkimusaineiston graafiseen ja numeeriseen kuvaamiseen. Tämän jälkeen esitellään tilastollisen päättelyn periaatteita ja yksinkertaisimpia tilastollisia testejä yhtäsuuruuksiin ja riippuvuuksien väliseen tarkastelemiseen. Seuraavaksi käsitellään tilastollisen mallintamisen ja luokittelun perusteita, ja kirjan loppuosa keskittyy aineistoa tiivistävien tilastollisten monimuuttujamenetelmien käsittelemiseen.

Mitä paremmin ymmärrämme luonnon ja yhteiskunnan toimintaa, sitä monimutkaisempaa huomaamme sen olevan. Lisäksi mittalaitteiden ja mittaustapojen kehittyminen lisää saatavilla olevan datan määrää tehden aineistojen analysoimisesta vuosi vuodelta vaativampaa. Onneksi tilastolliset ohjelmistot pysyvät mukana kehityksessä. Nykypäivänä yksi tilastollinen ohjelmisto on merkittävästi kilpailijoitaan parempi. Avoimeen lähdekoodiin perustuva, ilmainen ja valtavan kansainvälisen verkoston kehittämä ja tukema R (<https://cran.r-project.org>) on 2000-luvulta alkaen ollut tilastotieteen *lingua franca* -yleiskieli jota tilastotieteilijät ympäri maailmaa puhuvat äidinkielenään, mutta jota myös ”muunkieliset” toisten alojen tutkijat ymmärtävät sujuvasti. Tämän vuoksi tämän kirjan analyysiesimerkit perustuvat R-koodin käyttöön. Vaikka R perustuu komentokieleen ja tilastolliseen ohjelmointiin, ei kannata pelästyä. Ohjelmointia voi lähestyä kuten aloitteleva ruuanlaittaja keittokirjaa. Hyvästä keittokirjasta löytyy helppotajuiset reseptit, joita noudattamalla aloittelijakin onnistuu ruoanlaitossa, ja voi vähitellen alkaa muuntelemaan ja parantelemaan alkuperäisiä reseptejä. Samalla tavalla ohjelmointiakin voi opetella aluksi kopiaimalla ja

muuntelemalla olemassa olevaa ohjelmakoodia. Jos et vielä tätä kirjaa lukiessasi osaa ohjelmoida, niin nyt on erinomainen syy aloittaa opiskelu! Harvasta asiasta on niin paljon hyötyä kaikilla elämän alueilla kuin ohjelmointitaidosta, koska se opettaa ajattelemaan algoritmeja, tiedon hallintaa sekä prosessien toteuttamista niiden automatisoimisen näkökulmasta.

Tilastollisten menetelmien käyttäminen vaatii usein improvisointia. Jokainen tutkimusaineisto voidaan analysoida lukuisilla erilaisilla tilastollisilla menetelmillä, eikä mikään näistä ole välttämättä ainoa oikea tai absoluuttisesti paras. Samoin luonto ajautuu usein törmäyskurssille tilastotieteen teorian kanssa. Tässä kirjassa on koetettu lähestyä tutkimusaineiston analysoimista mahdollisimman realistisesta näkökulmasta. Siistien ja ongelmattomien aineistoesimerkkien sijaan kirjassa käytetään todellisia tutkimusaineistoja, jotka eivät läheskään aina ole täydellisiä. Aineistoja ja esimerkkejä on kerätty useilta tieteenaloilta. Kiitän lämpimästi kaikkia omia aineistojaan esimerkkikäyttöön luovuttaneita tutkijoita, joiden täydellinen lista löytyy kirjan lopusta. Kaikki kirjassa käytetyt esimerkkiaineistot ja -koodit voi ladata omalle koneelle osoitteesta emotion.utu.fi/tilastotiede.

Tutkimus ja tilastollinen analyysi on luovaa työtä siinä missä soittaminen, kirjoittaminen tai maalaaminenkin. Tilastotiede on tutkijan työkalu samoin kuin asteikot ja harmonia ovat muusikon työkaluja, tai väripaletti maalarin työväline. Ilman työkalujen hyvää hallintaa ei itse työstäkään tule mitään. Toisaalta myöskään pelkkä työkalujen hallinta ei riitä kovin pitkälle. Teknisesti ja harmonisesti täydellinen musisointi on puuduttavaa kuunneltavaa, jos siitä puuttuu musiikillinen idea. Samoin tilastotieteen näkökulmasta virheettömästi suoritettu tutkimus voi olla tyhjänpäivästä, jos tutkijalta puuttuu oivallus ja näkemys siitä, mikä maailmassa on kiinnostavaa ja merkityksellistä. Hyvä tutkija hallitseekin sekä oman tutkimusalansa sisällöt puhtaan teknisen tilastollisen osaamisen lisäksi. Yhdessä kirjassa on mahdotonta esittää kattavasti kaikkia mahdollisia tilastotieteen osa-alueita. Tässä kirjassa esiteltyjen menetelmien avulla voi kuitenkin hallita ja käsitellä jo hyvin monipuolisia tutkimusaineistoja. Tilastotieteen teoria ja menetelmät päivittyvät jatkuvasti, ja tämän kirjan tiedoilla selviää erinomaisesti myös tulevaisuudessa uusia tilastomenetelmiä opetellessa ja soveltaessa.

Turussa 8.6.2021

Tekijä

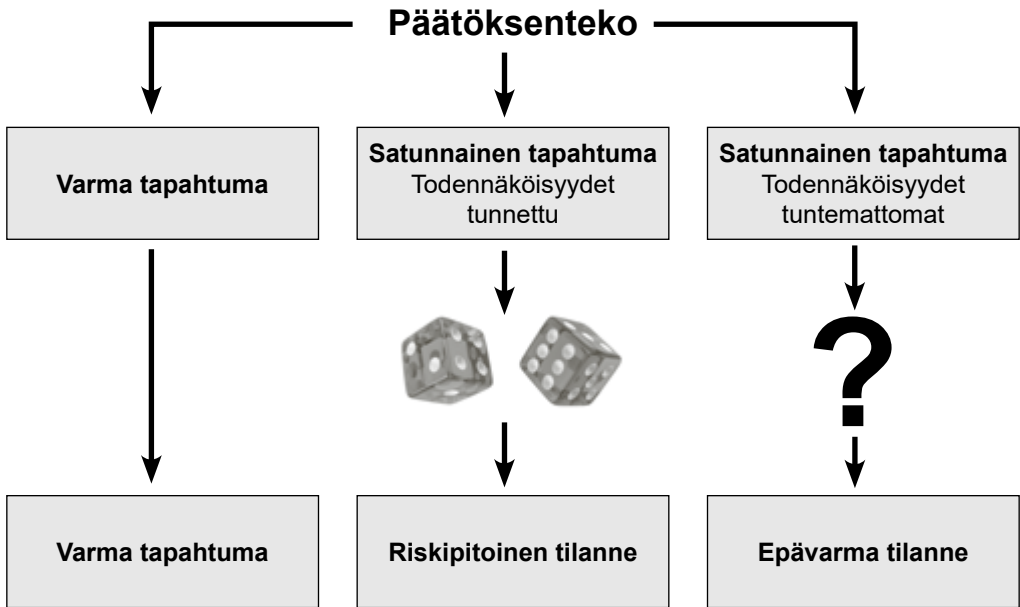
1. EPÄVARMUUDEN VÄHENTÄMINEN

Kaikki tutkimus perustuu tavalla tai toisella mittaustulosten numeeriseen tai tilastolliseen käsittelemiseen. Tilastollisten menetelmien tunteminen on tarpeellista paitsi tutkimusta tehdessä myös tutkimusraportteja lukiessa. Tieteellisten artikkelien lisäksi sanomalehdet, kirjat ja nettisivut ovat täynnä tilastollista tietoa – esimerkiksi talouden kehittymistä kuvaavia lukuja sekä työllisyyden kehitystä tai ilmaston lämpenemistä esittäviä tilastollisia kuvaajia. Tässä luvussa kerrotaan, kuinka tilastollisten menetelmien avulla voidaan tiivistää suuria tietomassoja ja kuinka niiden avulla voidaan tehdä luonnon ja yhteiskunnan ilmiöitä koskevia päätelmiä.

Järjestystä kaaoksen keskellä

Maailma on monimutkainen paikka. Ihmisaivot tekevät jatkuvasti parhaansa ennustaakseen ja ymmärtääkseen ympäristömme ja sosiaalisen pelikenttämme lainalaisuuksia, ja useimmiten ne toimivat erinomaisesti. Osaat varmasti sanoa kokemuksesi perusteella, ovatko miehet yleensä naisia pidempiä (ovat), kesäpäivät talvipäiviä lämpimämpiä (ovat), tai ovatko lapset aikuisia vahvempia (eivät ole). Toisaalta mitkään edellisistä väittämistä eivät pidä *täysin* varmasti paikkaansa. Tunnet varmasti monta miestä, jotka ovat lyhyempiä kuin jotkut tuntemasi naiset. Suomen kesäoinä lämpötila voi joskus laskea alemmas kuin poikkeuksellisen leutoina talvipäivinä, ja joillain lapsilla voi olla enemmän voimia kuin esimerkiksi huonokuntoisilla aikuisilla tai vanhuksilla. Kuinka voimme kuitenkin toimia ja tehdä erilaisia itseämme ja muita koskevia päätöksiä, jos emme voi olla täysin varmoja siitä, mitä päätöksistämme seuraa?

Päätöksenteko voidaan jakaa karkeasti kolmeen erilaiseen tilanteeseen (**Kuva 1.1**). Jos tiedämme täysin varmasti mitä päätöksistämme seuraa, päätös on turvallinen tehdä. Tiedämme esimerkiksi, että vapaasti palavien puiden synnyttämä lämpö ei riitä sulattamaan terästä, joten voimme valmistaa teräksestä turvallisen kiukaan saunaan. Toisaalta tiedämme, että sama palavien puiden lämpö, joka ei riitä sulattamaan terästä, aiheuttaa ihoomme kivuliaan palovamman, joten ymmärrämme olla koskematta kuumaan kiukaaseen. Toisessa ääripäässä olemme tuntemattoman tai hallitsemattoman ilmiön edessä, emmekä tiedä



Kuva 1.1. Päätöksenteko turvallisessa, riskipitoisessa ja epävarmassa tilanteessa.

lainkaan mitä päätöksestä seuraa. Esimerkiksi rahamarkkinoiden heiluessa rajusti sijoittaja ei voi välttämättä kuin arvailla, kannattaako hänen myydä osakkeitaan vai ostaa niitä lisää, koska ennakkotieto ei riitä tulevan kurssikehityksen ennustamiseen.

Tällainen satunnaisuus kuuluu ihmiselämään sekä kaikkeen mitä luonnossa tapahtuu. Satunnaisuus ei kuitenkaan tarkoita samaa kuin kaoottisuus. Vaikkemme voikaan täysin varmasti tietää mitä erilaisista päätöksistä seuraa, emme useimmiten joudu toimimaan täysin sokkona. Suurin osa tapahtumista onkin näiden kahden ääripään – täysin varman ja enemmän tai epävarman – välissä. Useimmat tapahtumat ovat nimittäin – toiset enemmän ja toiset vähemmän **ennustettavia**: tiedämme useimmiten, kuinka **todennäköisiä** eri lopputulokset ovat. Tiedämme esimerkiksi, että voimme lentokoneella matkustaessamme joutua onnettomuuteen ja kuolla. Mutta – koska tiedämme myös, että lento-onnettomuuteen liittyvä riski on häviävän pieni (noin yksi 11 miljoonasta), emme juurikaan murehdi lentomatkestamista. Vastaavasti tiedämme, että todennäköisyys voittaa suomalaisessa lotossa on vielä pienempi (noin yksi 15 miljoonasta), joten haaveillessamme lottovoitosta ymmärrämme, että unelmamme ei kovinkaan varmasti toteudu. Tupakoinnin taas tiedämme lisäävän keuhkosityövän riskiä niin paljon (noin 10–50 kertaa verrattuna siihen, ettemme tupakoi), että jätämme mielellämme savukkeet sytyttämättä.

Kaikissa edellisissä tapahtumissa molemmat lopputulokset ovat mahdollisia: Useimmilla kierroksilla joku ihminen voittaa lotossa ja suurin osa ainoastaan menettää rahansa, joka vuosi miljoonien turvallisten lentomatkojen lisäksi sattuu harvoja lento-onnettomuuksia, ja aina jotkut onnekkaat tupakoitsijat onnistuvat välttämään keuhkosityövän ankarasta sauhuttelustaan huolimatta. Vaikkemme tietäisikään mitä itsellemme tulee tapahtumaan, kun astumme koneeseen, täytämme lottokupongin tai jätämme savukkeet kauppaan,

todennäköisyyksien tunteminen auttaa meitä riskipitoisissa päätöksentekotilanteissa. Mutta kuinka voimme määritellä erilaisiin luonnon, yhteiskunnan ja talouden ilmiöihin liittyviä todennäköisyyksiä ja käyttää niitä päätöksenteon apuna?

Kuinka näkymätön saadaan näkyväksi?

Kolera oli vielä 1800-luvun puolivälissä vaarallinen tartuntatauti, jonka aiheuttajaa ei tunnettu. Teollisen vallankumouksen johdosta Lontoo oli nopeasti kasvava suurkaupunki, jossa erityisesti työläisten olot olivat tiiviin asutuksen ja huonon hygienian vuoksi surkeat. Suuri osa jätteistä laskettiin surutta Thames-jokeen, josta myös otettiin vettä kotitalouksien käyttöön. Koleran ja muiden tartuntatautiin tarttumismekanismeja ei vielä tunnettu. Tutkijat ymmärsivät tautien kulkeutuvan ihmisestä toiseen, mutta tutkijat olivat jakautuneet kahteen leiriin – osa uskoi koleran leviävän ilmaitse ”miasmojen”, myrkyllisten huuруjen välityksellä. Toiset tutkijat taas ajattelivat tauteja aiheuttavien mikrobien leviävän ulosteiden kautta vaatteisiin ja käyttöveteen. Vuonna 1853 Lontoon kolmannen suuren koleraepidemian aikaan brittiläinen lääkäri John Snow ryhtyi suorittamaan tilastollista tutkimusta kolera- tapauksista selvittääkseen pitääkö mikrobiteoria paikkansa. Hän vertaili koleratapausten esiintymistä kahdella Lontoon kaupunkialueella, jotka olivat muuten samanlaisia, mutta jotka saivat juomavetensä eri vesiyhtiöiltä. Snow huomasi, että koleratapauksia oli huomattavasti vähemmän alueella, jonne vesi toimitettiin Thamesin yläjuoksulta, minne jätevesiä ei juurikaan laskettu. Sen sijaan koleratapauksia oli paljon sellaisen vesiyhtiön jakelualueella, joka otti vetensä keskusta-alueelta, jossa siihen pääsi sekoittumaan viemäriverettä. Kesken tutkimuksen Snow’n kotiseudulla Lontoon Sohossa puhkesi yllättäen raju epidemia. Snow haastatteli asukkaita ja laati haastattelujen perusteella kartan sairastuneiden potilaiden osoitteista (Kuva 1.2). Hän havaitsi tautitapausten keskittyvän alueelle, jossa ihmiset ottivat vetensä Broad Streetin yleisestä kaivosta, johon pääsi valumaan vettä läheisestä suuresta likakaivosta. Snow taivutteli kaupunginvaltuuston poistamaan Broad Streetin saastuneen kaivon pumppukahvan, jotta epidemia saataisiin talttumaan.

Myöhemmin Snow esitti havaintonsa kuuluisassa kartassaan, jossa koleratapaukset on merkitty Sohon alueen karttaan pisteinä (Kuva 1.2). Pisteet keskittyvät selvästi Broad Streetin kaivon ympäristöön. Tämä tutkimussarja sai Snown vakuuttuneeksi siitä, että taudit voivat tarttua nesteissä olevien mikrobien mukana. Snow analysoi kaivon vesinäytteitä sekä kemiallisesti että mikroskoopin avulla, mutta ei pystynyt aukottomasti osoittamaan vedessä olevan taudinaiheuttajia. Sen sijaan hän päätteli tautitapausten esiintymisen ja kaivon sijainnin välisestä yhteydestä, että veden



Kuva 1.2. John Snow’n kartta koleratapausten esiintymisestä Lontoon Sohossa. Mitä lähempänä Broad Streetin kaivoa talot sijaittivat, sitä enemmän koleratapauksia havaittiin. Muiden samanlaisten kaivojen yhteydessä samanlaista tautikeskittymää ei havaita.

käyttäminen oli tautitapauksia yhdistävä tekijä: Broad Streetin pumpun lähistöllä sijaitse-
vissa taloissa koleratapauksia oli merkittävästi enemmän kuin muilla vastaavilla alueilla ja
samanlaisten kaivojen läheisyydessä. Myöhemmin 1800-luvun lopulla Louis Pasteurin ja
Robert Kochin tutkimukset todistivat tiettyjen tautien olevan elävien olentojen, mikrobien,
aiheuttamia. Snow'n tekemät tarkat **havainnot** tautitapausten ilmenemisen **lainalaisuuk-
sista** sekä niiden **esittäminen** karttana osoittivat kuitenkin **tilastollisesti** yhteyden likaisen
veden ja sairastumisen välisen yhteyden, ja tutkimusta pidetään eräänä ensimmäisenä
epidemiologisena eli tautien tartuntamekanismeja selvittävänä tutkimuksena. Karttansa
avulla Snow teki kirjaimellisesti näkymättömästä näkyvän – hän pystyi osoittamaan,
kuinka ihmissilmälle näkymättömien (ja tuolloin tuntemattomien!) kolerabakteerien leviä-
minen noudatti selkeitä lainalaisuuksia.

Tutkimustiedon tilastollinen käsittely

Nykypäivänä eräs tärkeimmistä keinoista epävarmuuden vähentämiseen ja lainalaisuuksien
havaitsemiseen on **tilastotiede**. Tilastotieteestä tulee helposti mieleen tilastotietojen me-
kaaninen tallentaminen ja käsitteleminen, ja tämä onkin yksi tilastotieteen tehtävistä.
Huolellisen tilastokirjanpidon vuoksi tiedämme esimerkiksi, kuka on pelaaja tehnyt NBA-ural-
laan eniten pisteitä (Kareem Abdul-Jabbar), kenellä rocklaulajalla on laajin tunnettu ääniala
(W. Axl Rosella) tai mikä on kaikkien aikojen myydyin musiikkialbumi (Michael Jacksonin
Thriller). Tämä on kuitenkin vain pieni osa tilastotieteen monentyyppisistä tehtävistä ja
sovelluksista. **Tilastotiede** on menetelmätiede, joka tutkii sitä, miten erilaisiin havaintoihin
ja mittauksiin perustuva tutkimus pitää suorittaa. Tilastotieteen tutkimusalueisiin kuuluu
mittaustulosten analysointiin, kuvaamiseen ja soveltamiseen liittyvien menetelmien kehit-
täminen ja päätöksentekoa koskevan epävarmuuden vähentäminen näiden menetelmien
avulla. Tilastotieteen menetelmiä tarvitaan kaikilla tieteenaloilla aina soveltavasta fysii-
kasta taloustieteeseen ja humanistisiin tieteisiin. Tilastotiede ei kuitenkaan ole mikään
muiden tieteenalojen aputiede, vaan oma itsenäinen tieteenalansa. Koska eri tieteenalat
tarvitsevat erilaisia tilastollisia analyysimenetelmiä, tilastotieteilijät tekevät usein tiivistä
yhteistyötä muiden alojen tutkijoiden kanssa. Monet nykypäivänä käytetyistä tilastomen-
etelmistä ovat syntyneet, kun uudet tutkimuskohteet ovat vaatineet uudenlaisten analyysi-
menetelmien kehittämistä.

Maailma on täynnä asioita, olioita ja tapahtumia. Ihmiset havainnoivat jatkuvasti ym-
päristöään ja tekevät tulkintoja ilmiöiden syistä, seurauksista ja lopputuloksista. Erilaisten
havaintojen tekeminen ja tulkitseminen kuuluu myös tieteelliseen tutkimukseen. Tutkijan
havainnointi poikkeaa kuitenkin merkittävästi arkihavainnoinnista. Tieteellisen tutkimuksen
tavoitteena on kerätä **järjestelmällisesti** tietoa luonnon toiminnasta ja muodostaa kerätyn
tiedon perusteella teorioita asioiden ja ilmiöiden välisistä yhteyksistä. Teoriat ovat **malleja**
tai yksinkertaistuksia todellisuudesta, jotka mahdollistavat tulevien tapahtumien ennusta-
misen. Muodostetut teoriat puolestaan ohjaavat tulevien tutkimusten suorittamista.
Vastaavanlaista mittaamisen, havaintojen tulkinnan ja teorianmuodostuksen menetelmä
voidaan soveltaa monissa asiantuntijatehtävissä – ratkaistavat ongelmat vain vaihtelevat
sovellusalueittain. Ensimmäiseksi ongelma muotoillaan, seuraavaksi kerätään ongelman-

ratkaisun kannalta tarpeellista tietoa, ja viimeiseksi hankitun tiedon perusteella tehdään tulkintoja ja pyritään muodostamaan vastaus ongelmaan.

Kaikissa mittaustuloksissa esiintyy vaihtelua. Ihmisten verenpaine, tulehdusarvot tai aivojen koko vaihtelevat tutkittavasta ihmisestä toiseen. Osa vaihtelusta on satunnaista. Pienet muutokset ihmisen vireystilassa tai elimistön perustoiminnoissa voivat aiheuttaa hetkittäisiä, terveyden kannalta merkityksettömiä muutoksia verenpaineessa. Osa vaihtelusta on puolestaan systemaattista. Tupakointi, vähäinen liikunta ja ylipaino ovat kaikki järjestelmällisesti yhteydessä kohonneeseen verenpaineeseen ja sitä kautta sairastumisriskiin. Tilastollisten menetelmien yksi tarkoitus on tällaisen – satunnaisen ja systemaattisen vaihtelun – kuvaileminen ja tarkasteleminen. Mutta kuinka voimme erottaa satunnaisen vaihtelun systemaattisesta vaihtelusta? Monet tutkimusaineistot ovat niin suuria, ettemme pysty hahmottamaan niitä paljaalla silmällä lainkaan. Ja vaikka joskus voisimmekin, on pelkkien käsittelemättömien mittaustulosten ymmärtäminen useimmiten mahdotonta. Tällaisten ongelmien ratkaisemiseen tarvitaan tilastotiedettä ja erilaisia tilastollisia menetelmiä. Kaikkien tilastollisten menetelmien perustavoite on sama – ilmiötä koskevan epävarmuuden vähentäminen. Tilastotiede käyttää tähän viidenlaisia menetelmiä:

1. **Aineiston rakenteen yksinkertaistaminen ja kuvaileminen.** Mittalaitteilla kerätyt aineistot ovat usein liian monimutkaisia suoraan ymmärrettäväksi. Aineiston ymmärtämiseksi on usein välttämätöntä tiivistää aineistoa ja yksinkertaistaa sen rakennetta. Tämän vuoksi aineistojen yleisiä ominaisuuksia voidaan esittää yksinkertaistetusti **tilastollisten tunnuslukujen** ja erilaisten **kuvaajien** avulla.
2. **Muuttujien välisten yhteyksien tutkiminen.** Suuri osa tutkimuksista keskittyy tavalla tai toisella erilaisten ilmiöiden välisten säännönmukaisten yhteyksien tutkimukseen. Erilaiset muuttujien välisiä riippuvuusuhteita arvioivat menetelmät ovat kenties tilastotieteen käytetyin menetelmäjoukko.
3. **Ilmiötä koskevien hypoteesien testaaminen ja ilmiöiden yleistyvyyden arvioiminen.** Maailma on täynnä ihmisiä, eläimiä ja molekyyliä. Emme voi mitata millään niitä kaikkia, vaan tutkimuksen mittaukset voidaan suorittaa vain pienelle osajoukolle. Hypoteesien testaamisen avulla voimme arvioida, kuinka todennäköisesti mittaustuloksissamme esiintyvä ilmiö **toistuu** ja **yleistyy** myös muihin tilanteisiin.
4. **Lajittelu ja ryhmittely.** Monet tutkimusongelmat liittyvät asioiden tai mittaustulosten muodostamien luokkien määrittämiseen – esimerkiksi siihen, millaisia eläinlajeja tai ihmisryhmiä aineistomme sisältää. Ryhmittelyyn perustuvien tilastollisten menetelmien avulla voimme yrittää löytää mittaustulosten sisältämiä **luokkarakenteita**.
5. **Ennusteiden laatiminen ja mallintaminen.** Pelkkä nykyhetken kuvaileminen on tärkeää, mutta useimmiten haluamme tutkimustulosten avulla myös ennustaa tulevaisuutta. Tilastollisten menetelmien avulla voidaan muodostaa erilaisia **malleja** joita voidaan käyttää tulevaisuuden tapahtumien ennakoimiseen.

Aineiston kuvaileminen ja yksinkertaistaminen

Ihmisen kyky käsittää suuria tietomääriä on rajallinen. Katso esimerkiksi **kuvan 1.3A** numerojoukkoa. Sen sisältämästä tiedosta on lähes mahdotonta saada minkäänlaista selkoa, vaikka numerot sisältävätkin merkityksellistä tietoa – numeromatriisi esittää nimittäin pientä osaa **kuvasta 1.3B**, tarkalleen ottaen tytön oikeaa silmää (**1.3C**). Kun vertaat matriisia ja sen alla olevaa silmän kuvaa, pystyt huomaamaan yhtäläisyyksiä – niissä kohdissa, missä on silmän kovakalvo, on suuria numeroita, koska kuva on näistä kohdin kirkas. Hieman pienemmät numerot puolestaan kuvaavat kasvojen ihoa, joka on vähemmän kirkas kuin silmän kovakalvo. Kaikkein pienimmät numerot puolestaan kuvaavat tummia kulmakarvoja ja silmän pupillia. Itse asiassa digitaalikamera tallentaa kuvan samaan tapaan pikselikohtaisina numeroarvoina. Vaikka kuvassa ja matriisissa on täsmälleen yhtä paljon tietoa, niiden käsittely vaatii aivan erilaista ponnistelua – suurten lukujoukkojen käsitteleminen on ihmiselle vaikeaa myös silloin, kun tiedämme etukäteen mitä numerot kuvaavat. Ja useimmiten tutkimuksessa ei etukäteen tiedetä, mitä tutkimukseen liittyvät mittaustulokset pitävät sisällään. Tämän vuoksi tarvitsemme samanlaisia työkaluja kuin digitaalisen kuvan sisältämän numerotiedon muuttaminen kuvaksi. Niiden avulla voimme tehdä ihmisilmälle alun perin **näkymättömistä ilmiöistä näkyviä**. Ja kuten valokuvaa katsottaessa, myös tilastolliset menetelmät auttavat meitä katsomaan dataa **riittävän kaukaa**. Näemme **kuvan 1.3B** tarkkana, koska emme katso sitä liian läheltä. Sen sijaan **kuvassa 1.3C** kuvan pikselit erottuvat, ja silmää on hankala erottaa. Mutta jos siirret kirjaa riittävän kauas, myös silmän kuva on helppo hahmottaa. Tilastollinen kuvaileminen toimii samaan tapaan – sen avulla voimme ottaa tarpeeksi etäisyyttä tutkimusaineiston numerotietoon, jotta voisimme kirjaimellisesti nähdä metsän puilta.

Suuri osa tilastollisen tutkimuksen tekemisestä liittyy mittaustulosten **kuvailemiseen** mahdollisimman helposti ymmärrettävällä tavalla. Kun pystymme ymmärtämään mittaustuloksien sisältämät ilmiöt paremmin, pystymme tekemään parempia tulevaisuutta koskevia päätöksiä. Esimerkiksi köyhyyden poistaminen on tärkeä ihmisten hyvinvointia edistävä tekijä, mutta miten köyhyydelle on maailmassa käynyt? Utisista saamme jatkuvasti kuulla tuloeroista, eriarvoisuudesta ja luonnonkatastrofeista. Tällainen tietotulva saa meidät helposti **kokemaan**, että asiat ovat menossa maailmassa huonompaan suuntaan. Tutkija ei kuitenkaan luota kokemukseen vaan **järjestelmällisesti kerättyyn tietoon**. Talous toimii hyvinvoinnin polttoaineena, joten Maailmanpankki (<http://www.worldbank.org/>) kerää jatkuvasti esimerkiksi köyhyyteen liittyvää maakohtaista tietoa, minkä perusteella voimme helposti arvioida, mitä köyhyydelle on todellisuudessa tapahtunut. Emme millään pystyisi hahmottamaan aineistoa, jossa olisi jok’ikisen maapallon asukkaankin tulot. Niinpä onkin käytännöllistä, että Maailmanpankki on jo käyttänyt tilastollisia menetelmiä **tiivistääkseen** köyhyyttä koskevaa tietoa. **Taulukossa 1.1** on esitetty absoluuttisen köyhyysrajan (alle \$2 päivässä) alla elävien ihmisten lukumäärä vuositasolla muutamalla maapallon köyhimmällä alueella. Tällaisessa taulukossa on miljardien havaintojen sijaan vain muutamia kymmeniä havaintoa. Taulukkoa silmäilemällä vaikuttaisi siltä, että näillä alueilla lukumääräisesti köyhiä on ollut eniten Itä- ja Etelä-Aasiassa ja Saharan eteläpuolisessa Afrikassa. Yhteenlaskettujen köyhyydessä elävien ihmisten määrän perusteella näyttäisi siltä, että köyhyys on

A) Kuvan osa numeroina

```

100 106 202 200 187 185 187 186 197 189 106 102 100 104 104 104 102 102 106 102 104 102 101 103 100 104
100 100 107 100 107 105 100 102 100 104 100 100 102 101 176 177 182 178 174 181 175 100 100 100 100 100
101 105 103 107 100 100 100 101 101 176 176 174 174 174 174 174 174 174 174 174 174 174 100 100 104 100
100 100 100 100 100 102 178 179 174 170 170 128 128 127 127 147 155 159 142 159 86 71 80 132 144 147 104
105 100 104 100 178 178 181 104 100 100 140 104 100 100 100 111 110 110 110 100 100 100 100 100 100 100
107 100 102 101 102 179 179 100 100 128 128 127 127 147 155 159 142 159 86 71 80 132 144 147 104
101 177 178 180 178 180 100 102 132 140 107 178 178 102 147 156 127 121 30 40 47 94 114 132 104
102 178 180 174 148 103 100 104 100 100 170 120 96 70 46 45 57 47 44 43 45 51 82 134 142
173 107 106 134 130 100 100 102 102 106 81 91 97 86 91 112 148 100 47 77 85 43 87 85 126
178 100 140 100 100 100 101 100 95 97 110 108 172 107 208 207 201 147 65 79 94 43 86 104 133
179 107 173 179 130 100 94 106 101 201 201 107 212 240 240 240 240 240 240 240 240 240 240 240 240 240
179 108 103 100 86 86 101 212 240 247 240 228 227 237 252 251 153 76 82 87 116 122 112 130
101 108 140 100 101 100 177 180 200 204 200 240 208 207 200 200 202 147 84 76 101 127 107 145 140
105 107 174 178 177 175 106 100 100 107 207 208 207 223 227 228 221 120 81 83 102 124 176 100 100
100 100 100 177 178 101 103 100 100 172 103 144 143 100 101 203 212 120 79 103 87 110 100 179 102
104 100 100 100 100 100 100 100 100 100 100 170 103 171 100 103 101 100 103 101 100 100 110 102 100 100
100 101 100 100 100 100 101 100 100 100 100 100 178 178 100 100 100 100 100 100 100 100 110 100 178
100 100 104 101 102 100 100 102 101 100 100 178 178 100 100 100 100 100 100 100 100 100 100 100 178 100
100 107 106 100 104 107 100 100 100 100 100 200 100 100 100 100 100 100 100 100 100 100 100 100 100 100
105 107 200 100 100 100 200 100 100 100 100 200 200 200 200 200 200 200 200 200 200 200 100 100 100 178 172
    
```

B) Alkuperäinen kuva



C) Kuvan osa

```

100 106 202 200 187 185 187 186 197 189 106 102 100 104 104 104 102 102 106 102 104 102 101 103 100 104
100 100 107 100 107 105 100 102 100 104 100 100 102 101 176 177 182 178 174 181 175 100 100 100 100 100
101 105 103 107 100 100 100 101 101 176 176 174 174 174 174 174 174 174 174 174 174 174 100 100 104 100
100 100 100 100 100 102 178 179 174 170 170 128 128 127 127 147 155 159 142 159 86 71 80 132 144 147 104
105 100 104 100 178 178 181 104 100 100 140 104 100 100 100 111 110 110 110 100 100 100 100 100 100 100
107 100 102 101 102 179 179 100 100 128 128 127 127 147 155 159 142 159 86 71 80 132 144 147 104
101 177 178 180 178 180 100 102 132 140 107 178 178 102 147 156 127 121 30 40 47 94 114 132 104
102 178 180 174 148 103 100 104 100 100 170 120 96 70 46 45 57 47 44 43 45 51 82 134 142
173 107 106 134 130 100 100 102 102 106 81 91 97 86 91 112 148 100 47 77 85 43 87 85 126
178 100 140 100 100 100 101 100 95 97 110 108 172 107 208 207 201 147 65 79 94 43 86 104 133
179 107 173 179 130 100 94 106 101 201 201 107 212 240 240 240 240 240 240 240 240 240 240 240 240 240
179 108 103 100 86 86 101 212 240 247 240 228 227 237 252 251 153 76 82 87 116 122 112 130
101 108 140 100 101 100 177 180 200 204 200 240 208 207 200 200 202 147 84 76 101 127 107 145 140
105 107 174 178 177 175 106 100 100 107 207 208 207 223 227 228 221 120 81 83 102 124 176 100 100
100 100 100 177 178 101 103 100 100 172 103 144 143 100 101 203 212 120 79 103 87 110 100 179 102
104 100 100 100 100 100 100 100 100 100 100 170 103 171 100 103 101 100 103 101 100 100 110 102 100 100
100 101 100 100 100 100 101 100 100 100 100 100 178 178 100 100 100 100 100 100 100 100 110 100 178
100 100 104 101 102 100 100 102 101 100 100 178 178 100 100 100 100 100 100 100 100 100 100 100 100 178 100
100 107 106 100 104 107 100 100 100 100 100 200 100 100 100 100 100 100 100 100 100 100 100 100 100 100
105 107 200 100 100 100 200 100 100 100 100 200 200 200 200 200 200 200 200 200 200 200 100 100 100 178 172
    
```

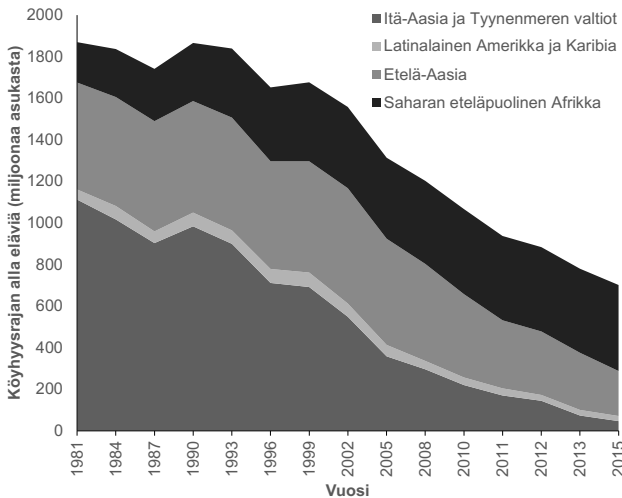
Kuva 1.3. Kuvan hahmottaminen numerojoukkona ja pikselien kirkkautena.

kokonaisuudessaan vähentynyt merkittävästi – seurantajaksolla yli **miljardi (!)** ihmistä on noussut pois absoluuttisesta köyhyydestä! Tällä mittarilla arvioituna ihmiskunnan kehitys ollut erittäin myönteistä.

Tietenkin taulukon tutkiminen on edelleen hankalaa – lukujen keskinäinen vertaileminen ja niiden suuruusluokan arvioiminen vaatii melkoista ponnistelua. Tämän vuoksi tilastollisen tiivistämisen ja kuvaamisen apuna käytetään usein **tilastollista grafiikkaa**. Ihmisivät ovat paljon taitavampia käsittelemään pinta-aloja, etäisyyksiä ja värejä kuin numeroita. **Taulukossa 1.1.** esitetty aineisto on helpompi hahmottaa, jos se esitetään graafisessa muodossa **kuvaajana (Kuva 1.4)**. Tässä **aluekuvaajassa** x-akselilla on kuvattu aika ja y-akselilla äärimmäisessä köyhyydessä elävien ihmisten määrä. Eri maantieteelliset alueet on esitetty erivärisinä monikulmioina, joidenka korkeus jokaisessa kuvaajan kohdassa kertoo köyhyyden määrän kyseisellä alueella. Monikulmiopinon yhteenlaskettu korkeus puolestaan kertoo köyhien kokonaismäärän. Tästä kuvaajasta näkyy selvästi, että köyhyys on vähentynyt tasaista tahtia ja että väheneminen on ollut kaikkein nopeinta Aasiassa. Sen sijaan Saharan eteläpuolinen Afrika vaikuttaisi pudonneen kehityksestä – siellä köyhyys on itse asiassa lisääntynyt hieman tarkastelujakson aikana.

Taulukko 1.1. Äärimmäisessä köyhyydessä elävien ihmisten lukumäärä (miljoonaa asukasta) kaikkein köyhimmillä alueilla.

Vuosi	Itä-Aasia ja Tyynenmeren valtiot	Latinalainen Amerikka ja Karibia	Etelä-Aasia	Saharan eteläpuolinen Afrikka	Yhteensä
1981	1112	50	514	194	1869
1984	1017	65	524	230	1836
1987	903	56	530	251	1741
1990	984	65	536	280	1865
1993	899	65	542	332	1838
1996	711	67	518	356	1651
1999	692	69	534	381	1677
2002	549	63	555	391	1558
2005	359	55	510	389	1313
2008	296	40	467	399	1202
2010	221	36	401	409	1066
2011	170	34	328	406	938
2012	144	29	305	406	883
2013	73	28	274	404	779
2015	47	24	217	413	701



Kuva 1.4. Äärimmäisessä köyhyydessä elävien ihmisten lukumäärä (miljoonaa asukasta) kaikkein köyhimmillä alueilla vuosina 1981–2015 Maailmanpankin (<http://worldbank.org>) tilaston perusteella.

Pelkästään yhden tunnusluvun tarkasteleminen ei tietenkään riitä siihen, että vakuutuisimme maailman tilan olevan menossa parempaan suuntaan. Siksi onkin hyvä tutkia samaan tapaan myös muita kuin pelkkiä taloudellisia muuttujia. **Kuvaan 1.5** on piirretty lisää maailman tilaa kuvaavia indikaattoreita elämän eri osa-alueilta. On selvää, että luku-taidottomien ihmisten määrä on romahtanut 1940-luvun jälkeen (A); samoin lapsikuolleisuus

Tilastollisten menetelmien merkitys tutkimuksessa kasvaa jatkuvasti. Tilastotiedettä tarvitaan kaikilla aloilla tutkimusaineiston keräämiseen, kuvailemiseen ja mallintamiseen sekä riskien arvioimiseen ja ennusteiden laatimiseen.

Teoksessa esitellään kattavasti keskeisimmät tilastolliset analyysimenetelmät alkaen aineiston keräämisestä ja kuvailemisesta sekä perusanalyyseista aina tilastollisiin monimuuttujamenetelmiin ja tilastolliseen luokitteluun asti.

Tilastollisten menetelmien teoria esitetään havainnollisesti, joten aloittelijakin pääsee analysoimaan erilaisia aineistoja itsenäisesti. Havainnolliset laskuesimerkit helpottavat asioiden omaksumista. Monipuoliset ohjelmistoesimerkit perustuvat kattavaan ja ilmaiseen R-ohjelmistoon, ja ne sisältävät selkeät esimerkkikoodit sekä ohjeet tulosteiden tulkinnasta.

Teos soveltuu yliopistojen ja ammattikorkeakoulujen tutkimuksen teon ja tilastollisten menetelmien perusoppikirjaksi lääketieteeseen, tekniikan, yhteiskunta- ja taloustieteiden sekä käyttäytymis- ja luonnontieteiden aloilla. Kattavuutensa ja monipuolisten esimerkkiensä ansiosta kirja sopii erinomaisesti myös varttuneemman tutkijan hakuteokseksi.

**Kattava ja ajantasainen
perusteos tilastollisten
menetelmien sovelluksista
opiskelijoille ja tutkijoille.**

Lauri Nummenmaa (FT) on maamme johtavia aivotutkijoita ja tilastollisen analyysin asiantuntijoita. Hän toimii lääketieteellisen kuvantamisen ja mallintamisen professorina Turun yliopistossa, jossa Nummenmaan ryhmä tutkii ja mallintaa aivojen toimintaa molekyyalitasolta alkaen.



9 789520 401382

www.tammi.fi

31

ISBN 978-952-04-0138-2