Surveymetodiikka

Aineiston kokoamisesta puhdistamisen kautta analyysiin

Seppo Laaksonen
Helsingin yliopisto
Sosiaalitieteiden laitos
Yhteiskuntatilastotiede
2014

Huipulle on monta reittiä.

Kurssin luonne

Luentoja on kuusi kertaa:

- 3 Viikkoa + Viikon tauko + 3 Viikkoa

Jos mahdollista koko 3 h yhtenä rupeamana mutta jos on tarve niin lyhyt tauko.

Harjoitukset mikroluokassa
Joko torstaisin klo 12-14 tai maanantaisin klo 16-18
Alkavat heti. Taukoviikkona ei ole harjoituksia, mutta jos olet jäljessä, toivon sinun niitä tekevän.
Harjoitustyöt raportoidaan ensimmäisissä harjoituksissa annettavan mallin mukaan. Niitä kannattaa tehdä mahdollisimman paljon harjoitusten aikana koska silloin voidaan antaa melkeinpä vastauksetkin.
Tentti viikon kuluttua viimeisestä luentokerrasta, myöhemmin yleistentissä. Kaikkiin ilmoittaudutaan etukäteen Weboodin kautta.
Toivon leppoisaa ja keskustelevaa kurssia.

Materiaaleista:

Nettikirjani kattaa melkein kaiken mitä vaaditaan tentissä.
Imuroi se siis
Joitakin lisäkkeitä tulee mukaan.
Käytän osin kirjaa suoraan mutta osin PPT-kalvoja joista osa on kirjasta, osa jokin lisäke.

Harjoituksissa on kahdenlaisia kansainvälisiä datoja:

- European Social Survey (ESS), valittavana nyt jo kuudes kierros, Round 6
  Tämä imuroidaan itse ensimmäisellä kerralla ohjatusti eli kannattaa olla paikalla.
- OECD:n Koulusaavutustutkimus PISA, uusin vuodelta 2012 (julkisuuteen 2013 ja 2014).
  Annan tästä itse imuroimani version myöhemmin.

Voidaan sivuta muutakin. Kannattaa katsoa kesäkuun 2014 Yhteiskuntapolitiikka –lehteä (http://www.thl.fi/fi/ajankohtaista/lehdet/yhteiskuntapolitiikka/arkisto/32014) mistä löytyy artikkelini joka kritisoi ja analysoi Tatu Vanhasen (Matin isä) artikkelia:
Tatu Vanhanen: Kansallisen älykkyysosamäärän merkitys PISA-tulosten vaihtelun selittäjänä
Seppo Laaksonen: Makroaineisto on huono yksilötason johtopäätöksiin

Kirjan sisällys	Kurssilla alustavasti
1. Mitä tarkoitan surveyllä	Katsotaan läpi, kannattaa palata
2. Surveyn vaiheet	Katso läpi mutta palaa aika ajoin.
3. Tiedonkeruu ja lomakesuunnittelu	Toisen kerran keskeinen asia
4. Survey-aineiston peruskäsitteistö	Ekakerralla ja myöhemmin
5. Otantamenetelmät	Kolmas kerta
6. Otantamenetelmien tekniikkaa	Kolmas ja neljäs kerta
7. Yksinkertaista estimointia ja otanta-asetelmavaikutus	Neljäs kerta
8. Otoskoko	Neljäs kerta
9. Puuttuneisuus	Viides kerta
10. Tilastollinen Editointi	Kuudes kerta, lyhyesti
11. Uudelleenpainotus	Kuudes kerta, lyhyesti
12. Imputointimenetelmät	Kuudes kerta, lyhyesti
13. Puhdistettu surveyaineisto	Kuudes, loppuyhteenveto

Liitteet

• Surveyesimerkkiliite	Uhritutkimus ekakerralla
• Epävarmuus tilastotieteessä liite	Sopivassa välissä
• Asteikko- ja muunnosliite	Lue itse, osa ekakerralla
• Lomakeliite	Luvun 3 oheen
• Analyysiliite	Viidennen kerran pääasia
• Sanastoliite	Katso sanoja ja mistä löytyy ja opi

Esimerkkisurveyt
- Harjoituksissa European Social Survey (ESS) ja OECD:n koulusaavutustutkimus PISA. Niiden selostukset tulevat harjoitusten yhteydessä. Nyt selostan käyttäen myös kurssin käsitteitä isoa tutkimushanketta josta on juuri ilmestymässä (oikoluettu siis) merkittävässä amerikkalaisessa lehdessä raporttimme:

Laaksonen, Seppo and Heiskanen, Markku (2014). Comparison of three modes for a victimization survey. *Journal of Survey Statistics and Methodology*. Oxford University Press.

Nyt kutsun tätä Uhritutkimukseksi, mutta taustalla on Eurostatin ja Oikeusministeriön rahoittama laaja hanke jossa testattiin erityisesti tiedonkeruuvälinettä turvallisuutta koskevassa kyselyssä. Hankkeessa mukana olivat Heuni (Kauko Aromaa, Markku Heiskanen, Elina Ruuskanen), Tilastokeskuksen haastattelijaorganisaatio (Hannu Virtanen ym.) sekä Helsingin yliopisto (Seppo Laaksonen ja myöhemmin graduntekijät Jenni Nikula, Marjukka Vartiainen ja Antti Pelanteri).

Tätä ennen olin Heunin kanssa vuoden hankkeessa jossa tehtiin tausta-analyysi koko kyselylle. Tällöin vuosina 2007-2008 vierailimme myös useissa maissa joissa vastaavia kyselyitä oli toteutettu ja laadimme seikkaperäisen raportin Eurostatille; sen liitteenä on laaja kysymyslomake. Piloteissa tutkitaan myös lomakkeen toimivuutta, ei toki kaikkien kysymysten osalta.

Suomen pilotin otoksen tein syyskuussa 2009, tiedot kerättiin sen jälkeen tammikuun 2010 puoliväliin mennessä, loppuraportti Eurostatille valmistui helmikuussa 2010. Jatkotutkimuksia on tehty sen jälkeen. Ajattelen, että se on nyt valmis mutta varmaan tulee kysymyksiä ja kommentteja pitkän aikaa.

Uhritutkimuksessa on kysymyksiä turvallisuuden tunteesta erilaisissa oloissa mutta erityisen paljon uhriksi joutumisesta. Tämä voi koskea varkautta, vahingontekoa, väkivaltaa ml. seksiä, myös uudet ongelmat kuten identiteetin varastaminen ja sen hyväksi käyttö ovat esillä. Voidaan yleisesti sanoa, että kysymykset ovat monesti aika rankkoja sekä kyselijälle että vastaajalle. Tästä syystä on odotettavissa että keruuväline voi vaikuttaa tuloksiin.

Me tutkimme täysin riippumattomasti kolmea keruutapaa, puhelinkyselyä, käyntihaastattelua ja nettiä. Kiinnostavaa oli myös nähdä kuinka paljon kuhunkin vastataan. Kerron joitakin tuloksia tässä mutta kirjassa on lisää. Toki meillä oli otantaa suunniteltaessa jotkin arviot koska halusimme riittävän paljon vastauksia jotta voimme verrata tuloksia. Myös haastattelijoiden asenteita ja kokemuksia tutkimme heille kohdistetulla kyselyllä.

Tämä tutkimus ei ollut vain pilotti vaan suomalaisella lisärahalla poimittiin miehistä suhteellisesti isompi otos. Näin saadaan uusia tuloksia miesten uhriksi joutumisesta, erityisesti heidän kokemastaan väkivallasta. Heunista saat seuraavan raportin käyttöösi.
Heiskanen, M. & Ruuskanen, E. (2011): *Men's experiences of violence in Finland 2009*. Publication Series No. 71. Helsinki: HEUNI (www.heuni.fi).

Seuraavassa selostan lyhyesti tämän surveyn vaiheet ja muutaman keskeisen tuloksen.

Suunnitteluvaihe Eurostatille jossa esitimme sen että haluamme verrata kolmea keruuvälinettä, tavallista käyntihaastattelua, puhelintiedustelua ja uutta välinettä nettiä. Tilastokeskukselle netti oli tässä ensimmäinen vakava alan kysely. Suunnitteluvaihe sisälsi luonnollisesti myös rahoitussuunnitelman eikä summa ollut pieni. Olin yllättynyt kun kaikki raha myönnettiin ja lisäksi saimme kotimaista rahoitusta. Siksi ei aluksi tuntunut työn tekeminen yhtään hankalalta, pikemminkin päinvastoin.

Tilastokeskuksen haastatteluvastaavien kanssa pidettiin tietysti useita palavereita joissa yksityiskohtia hiottiin. Myöhemmin myös haastattelijat testasivat kysymyksiä. Markku kävi myös kurssillani kertomassa hankkeesta ja opiskelijat testasivat lomaketta. He ehdottivat useita muutoksia siihen ja useimmat niistä toteutettiin.

Otoksen tein itse käyttäen Väestörekisteriä johon minulle hankittiin oikeudet. Huomaa että näitä oikeuksia kontrolloidaan tarkasti. Minullakaan ei näitä oikeuksia pitkään ollut mutta myös tein pari muuta otosta samoilla oikeuksilla.

Otantaa varten kokeilin ensi kertaa Tilastokeskuksen historiassa ryppäitä eli pienehköjä alueita. Tämä antaa mahdollisuuden siihen että käyntihaastattelijoiden matkat ovat pienehköjä. Toisaalta se antaa mahdollisuuden tutkia uhriksi joutumisen kasautumista alueellisesti. Molempien hyödyt havaittiin, edellinen käyntihaastattelussa mutta jälkimmäinen kaikissa kyselyissä.

Ryppäät muodostin posti- ja kuntanumeroiden perusteella siten että niitä oli koko maassa yhteensä 600. Tämä on hieman intuitiomainen ratkaisu mutta ei ole huonoksi havaittu. Otokseen poimin kaksiasteotannalla 100 ryvästä, ja kunkin näiden sisältä valittiin satunnaisesti 15-79 –vuotiaita tutkittavia. Tämä brutto-otos jaettiin satunnaisesti kullekin keruutavalle eli keruutavan vertailumahdollisuuden pitäisi olla ihanteellinen.

Otoskoko vaihtelu budjetin puitteissa. Kallein keruutapa on käyntihaastattelu, puhelin sitä puolta halvempi mutta netti 10:s osa käynnistä. Koska oli tuntuma, että netti on vähiten suosittu, ja puhelin ehkä suosituin väline, brutto-otoskoko suunniteltiin niin, että kaikista saadaan riittävän iso vastaajajoukko.

Käyntihaastattelun otoskoko oli vajaa 800 ja vastaajia vain puolet tästä. Tämä oli pettymys mutta vastaajia oli kuitenkin riittävästi. Olisi ollut parempi saada enemmän. Puhelinvastaajia saatiin runsas 60 % noin 3000:sta, mikä määrä antaa melko tarkkoja tuloksia. Nettiin otettiin noin 4000 mutta vastaajia oli vain 25%. Vastaajamäärä on kuitenkin hyvä tarkahkojen tulosten saamiseksi.

Otoksen pohjalla oli 32 ositetta eli tavoiteperusjoukon osajoukkoa. Miehiä poimittiin nettiin ja puhelintiedusteluun suhteellisesti enemmän. Myös nuoria otettiin enemmän koska oli pelko saada heiltä riittävästi vastauksia. Eli siis ositteiden suhteelliset osuudet on hyvä miettiä etukäteen hyvin sen mukaan mitä tuloksia halutaan ja mistä ryhmistä riittävän tarkkoja.

Voin näin jälkeenpäin sanoa, että ajattelumme oli ihan ok eli ei pahoja pettymyksiä tullut. Suurin ongelma koski haastattelua jossa uutterat haastattelijat kohtasivat käynnillään ja puhelimessa ihmisiä joilta tiedustelivat ikäviä asioita.

Väkivaltaa ja seksuaalista häirintää koskevat kysymykset koettiin erityisen stressaaviksi, mutta auto- tai polkupyörävarkaudet tai pelot vähemmän. Haastattelijoille tehty kysely osoitti tämän ja varmasti vaikutti tuloksiinkin, ehkä eniten puhelinkyselyn tuloksiin. Tämä näkyi siinä, että haastattelu meni usein liian nopeasti eikä tuloksia voi pitää hyvinä. Käyntihaastattelun etuhan on siinä, että haastattelija voi auttaa, motivoida ja muuten valmistella haastateltavaa, joka vastasi kaikkein herkimpiin aiheisiin anonyymisti eli haastattelijan antaman tietokoneen muistiin. Periaate on sama kuin nettikyselyssä jossa kukaan ei näe mitä vastaat.

Ennen kuin tuloksia voitiin laskea, on muodostettava kullekin kolmelle aineistolle hyvät painot (otospainot). Aluksi lasketaan ne vastaajille olettaen että kussakin ositteessa vastaaminen on satunnaista. Myöhemmin näitä parannetaan siten että otetaan huomioon rekistereistä saatavat vastaajien ja vastaamattomien taustatiedot. Periaate on nostaa paino jos vastaaminen on huonoa ja päinvastoin. Metodi sinänsä on vaikea eikä kuulu yksityiskohdiltaan tälle kurssille mutta lopussa sen ideat esitetään.

Tämä kuvio antaa käsityksen siitä miten iän mukaan kussakin ryhmässä vastattiin. Mitä korkeammalla viiva on, sitä paremmin vastasi.

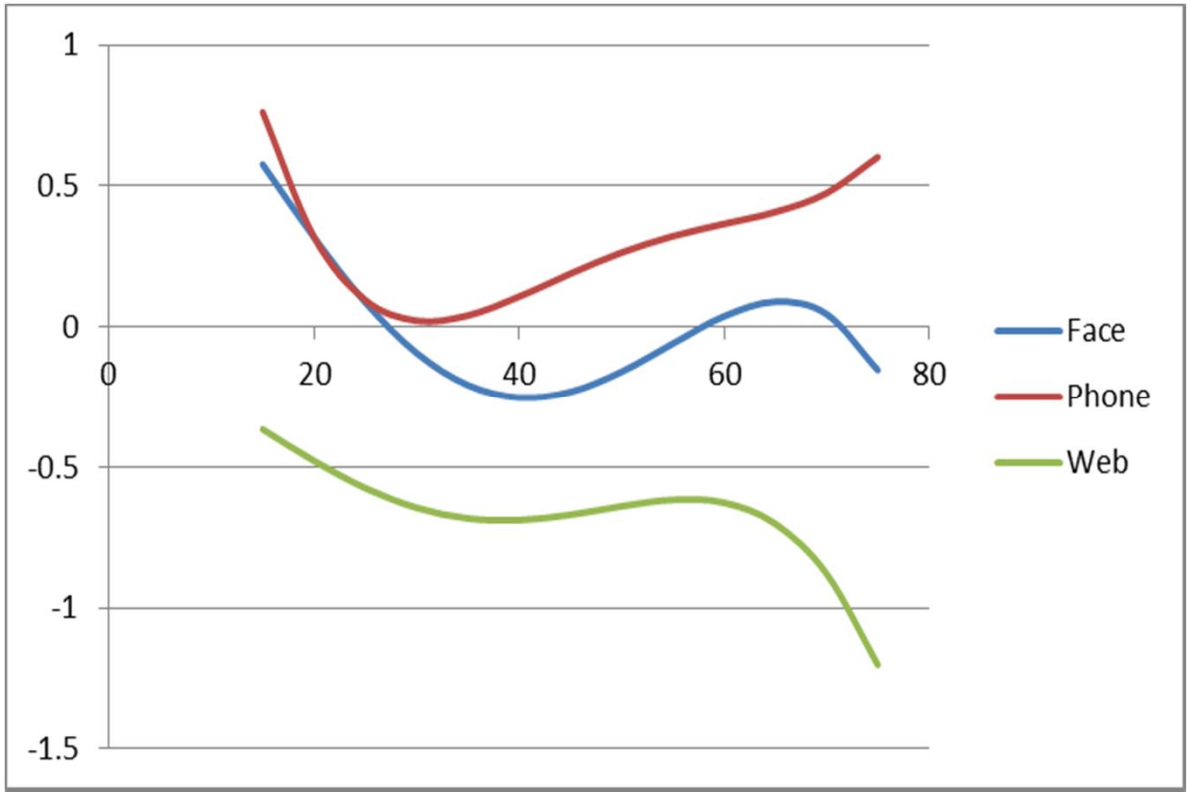


Fig. 3. Estimated effects of mode and age on response probabilities by age (x-axis), based on the probit model.
Älä välitä vaikket kaikkea tästä ymmärrä.

Lopuksi annan journalissa julkaistun tuloksen. Näet eroja mittareissa joiden minimi=0 ja maksimi =100.
Oikeimmat tulokset ovat ylemmät eli ’Adjusted’;
’Unweighted’ ovat vertailua varten. Voit katsoa onko eroja eli miten paljon yksinkertaisesti eli väärin lasketut luvut poikkeavat.

	Estimates (standard errors), adjusted above and unweighted below		
	Face-to-face	Telephone	Web
Fear	Interviewer-administered items	Interviewer-administered items	Self-administered
Feeling unsafe	18.2 (1.3) 17.7 (1.2)	17.0 (0.5) 15.6 (0.5)	20.6 (0.7) 19.6 (0.6)
Fear of burglary	24.5 (1.4) 24.7 (1.3)	21.0 (0.6) 20.7 (0.6)	26.5 (0.8) 25.6 (0.7)
Fear of assault	20.0 (1.3) 18.8 (1.1)	20.3 (0.6) 18.7 (0.5)	26.1 (0.8) 25.3 (0.7)
Fear of family or friends	33.5 (1.6) 32.6 (1.4)	32.2 (0.7) 31.5 (0.6)	37.6 (0.9) 36.6 (0.8)

Property crimes	Face-to-face	Telephone	Web
Theft of car, if car owner	1.8 (0.8) 1.5 (0.7)	2.5 (0.4) 2.2 (0,4)	3.8 (0.7) 3.6 (0.6)
Damage to car	9.9 (1.7) 11.3 (1.8)	10.0 (0.8) 9.6 (0.7)	15.5 (1.5) 15.2 (1.2)
Theft of bicycle	16.9 (2.1) 16.4 (2.0)	17.1 (1.0) 15.8 (0.8)	22.6 (1.6) 21.9 (1.3)
Burglary at free-time residence	6.8 (2.4) 6.1 (2.1)	8.5 (1.2) 7.8 (1.1)	9.0 (1.6) 9.4 (1.6)
Robbery	2.0 (1.0) 1.6 (0.7)	2.5 (0.4) 2.6 (0.4)	3.5 (0.7) 3.4 (0.6)
Theft, other personal property	9.5 (1.8) 8.7 (1.5)	9.4 (0.7) 9.0 (0.7)	12.1 (1.2) 11.8 (1.0)
Burglary at home	2.3 (0.8) 2.7 (0.9)	2.7 (0.4) 2.9 (0.4)	5.0 (0.8) 5.3 (0.7)
Fraud	8.5 (1.6) 8.7 (1.5)	9.6 (0.7) 9.3 (0.7)	10.3 (1.0) 11.4 (1.0)

Violent crimes	Face-to-face	Telephone	Web
Physical or sexual violence, last 5 years	10.3 (1.7) 10.4 (1.6)	9.3 (0.8) 8.4 (0.6)	15.1 (1.3) 15.0 (1.1)
Physical or sexual violence, last 12 months	3.5 (1.0) 3.6 (1.0)	3.9 (0.5) 3.5 (0.4)	6.2 (0.9) 6.4 (0.8)
	Self-adminis-tered (CASI)	Interviewer-administered items	Self-administered
Sexual harassment, since the age of 15	52.8 (2.8) 46.4 (2.6)	38.0 (1.2) 33.2 (1.0)	45.3 (1.8) 42.0 (1.5)
Sexual harassment, last 12 months	21.1 (2.4) 19.9 (2.1)	10.1 (0.8) 8.4 (0.6)	22.2 (1.5) 20.6 (1.3)
Violence by stranger, since the age of 15	36.2 (2.8) 36.9 (2.5)	30.3 (1.1) 31.3 (1.0)	37.2 (1.8) 38.9 (1.6)
Violence by stranger, last 12 months	11.0 (1.8) 10.1 (1.6)	4.4 (0.5) 4.6 (0.5)	9.0 (1.0) 9.5 (0.9)
Violence by partner, since the age of 15	16.4 (3.1) 14.3 (2.5)	9.4 (1.0) 9.7 (1.0)	20.1 (2.1) 20.2 (1.8)
Violence by partner, last 12 months	4.1 (1.6) 3.7 (2.5)	1.7 (0.5) 1.5 (0.4)	4.0 (1.0) 4.4 (0.9)

ASTEIKKO- ja MUUNNOSLIITE

Skaalat, muunnokset, suhteellisuus ja indikaattorit

On harvoin mahdollista käyttää alkuperäistä aineistoa sellaisenaan jatkoanalyysissä, vaan uusia muuttujia täytyy luoda vanhojen pohjalta. Periaatteessa näitä on kahta tyyppiä:

(i) Yksittäistä muuttujaa muunnetaan itsenäisesti jolloin se yleensä nimetään uudelleen.

(ii) Uusi muuttuja on kahden tai useamman alkuperäisen uusi kehitelmä.

1. Skaalat tai asteikot

Skaaloihin tai mittausmetriikoihin on perustasolla kaksi päälähestymistapaa, lähtien liikkeelle joko itse muuttujista ja niiden mittaamisesta tai muuttujien välisistä yhteyksistä. Tarkastelen asiaa ensin muuttujatyypeittäin etenevästi. Muuttujien käsittelyssä voi tulla harkittavaksi ainakin seuraavien skaalojen tai *mittausasteikkojen* käyttö:

(i) *Luokitteluskaala (nominaalinen)*, jota vain voidaan käyttää jos muuttuja itse on luokitteluasteikollinen kuten sukupuoli, kunta tai toimiala. Luokitteluskaalan 'alaskaaloja' on sankka määrä, ainakin eri nimiä.

- *Modaali* (modal)
- *Kategorinen* (*kategorisoida ylempi skaala*
- *Dikotominen* tai *kaksiarvoinen* tai *binäärinen tai dummy*

Paljon käytetty tällainen muuttuja on sellainen jossa on selvä vastakkainasettelu (komplementti) tyyliin 'On' vs 'Ei' jotka on kätevä koodata '1 vs 0' mutta voidaan toki skaalata muutenkin. Esimerkiksi kooditus 1='Kyllä' vs. –1='Ei' voi olla käyttökelpoinen tietyissä tilanteissa. Tällöin koodien keskikohta on =0. Nämä kaksiarvoiset muuttujat ovat paljon käytettyjä esimerkiksi kun erotellaan sairaita vs. terveitä, työttömiä vs. ei-työttömiä tai vastaajia vs. ei-vastaajia. Mutta 'kauniit vs. rohkeat' tai 'rikkaat vs. köyhät' eivät ole vastaavanlaisia.

(ii) Järjestysskaala, jolloin muuttuja itse on järjestysasteikollinen kuten koulutusaste tai kuntien tyypittely taajama-asteen mukaan (tähän ei ole tosin kiistatonta menetelmää). Myös standardiluokittelut kuten toimiala tai ammatti yritetään luokitella jonkin järjestysperiaatteen mukaan, esimerkiksi toimialan pääryhmät etenevät jalostusketjun ideaa pitkälle noudattaen eli alkutuotannosta jalostukseen ja edelleen kuljetukseen ja myyntiin. Loppupää ei tähän täysin istu, sinnehän on muun muassa sijoitettu yhteiskunnallisia toimintoja kuten yliopistot. Ammattiryhmäluokitukset taas lähtevät usein arvohierarkiasta, jossa useimmiten alkupäässä ovat johtajat ja erikoisasiantuntijat ja loppupäässä avustavissa tehtävissä toimivat. Mielipidetutkimuksissa käytetään laajasti järjestysasteikollisia skaaloja kuten 0, 1, …, 9, 10 tai 1, 2, 3, 4, 5; ne voivat olla myös tekstimuotoisia. (asteikko 1, 2, …, 10 on huono)

(iii) Välimatkaskaala, jossa välimatkat saadaan oikein mitattua, ja vastaavasti tietenkin järjestys. Näitä voidaan luokitella jos on tarpeen eli siirtyä järjestys- ja luokitteluasteikolle. Tyypillisiä tällaisia muuttujia ovat erilaiset muutokset, vaikkapa toimeentulotuen saajien määrällinen tai suhteellinen muutos kahden ajankohdan välillä. Luku voi olla positiivinen, negatiivinen tai nolla.

Huomaa, että kaksiarvoinen (0, 1)-muuttuja on myös järjestysasteikollinen ja voidaan käsitellä myös välimatka-asteikollisena. Siten jos esimerkiksi 0=mies ja 1=nainen, niin keskiarvo voidaan laskea mutta tulkinta tulee tehdä oikein. Jos esimerkiksi keskiarvo = 0,6 niin aineistossa on 60% naisia.
(iv) Suhdeskaala, jossa välimatkojen lisäksi voidaan suhteet mitata oikein. Tämä edellyttää absoluuttisen nollapisteen olemassaoloa. Toimeentulotuen saajien määrä tai keskimääräinen vastaanotettu markkamääräinen tuki, sekä kulutettu määrä jotain alkoholijuomaa juomatutkimuksissa ovat tyypillisiä esimerkkejä.

Välimatka- ja suhdeasteikon muuttujaa kutsutaan myös *jatkuvaksi* tai *kvantitatiiviseksi* eli *määrälliseksi* sekä myös *metriseksi.* Luokittelu- ja järjestysasteikon muuttujat taas ovat *kvalitatiivisia* tai *laadullisia*. Muut kuin luokitteluasteikon muuttuja ovat myös *monotonisia.*

(iii) Välimatkaskaala, jossa välimatkat saadaan oikein mitattua, ja vastaavasti tietenkin järjestys. Näitä voidaan luokitella jos on tarpeen eli siirtyä järjestys- ja luokitteluasteikolle. Tyypillisiä tällaisia muuttujia ovat erilaiset muutokset. Luku voi olla positiivinen, negatiivinen tai nolla.

Huomaa, että kaksiarvoinen (0, 1)-muuttuja on myös järjestysasteikollinen ja voidaan käsitellä myös välimatka-asteikollisena. Siten jos esimerkiksi 0=mies ja 1=nainen, niin keskiarvo voidaan laskea mutta tulkinta tulee tehdä oikein. Jos esimerkiksi keskiarvo = 0,6 niin aineistossa on 60% naisia.
(iv) Suhdeskaala, jossa välimatkojen lisäksi voidaan suhteet mitata oikein. Tämä edellyttää absoluuttisen nollapisteen olemassaoloa.

Välimatka- ja suhdeasteikon muuttujaa kutsutaan myös *jatkuvaksi* tai *kvantitatiiviseksi* eli *määrälliseksi* sekä myös *metriseksi.* Luokittelu- ja järjestysasteikon muuttujat taas ovat *kvalitatiivisia* tai *laadullisia*. Muut kuin luokitteluasteikon muuttuja ovat myös *monotonisia.*

Tietokoneohjelmistoissa on käytössä myös jako: *merkkimuotoiset* (character, string) vs. *numeeriset* (numeric). Käyttäjä voi valita missä muodossa muuttuja merkitään aineistoon. Joskus sama muuttuja voi olla hyvä sijoittaa aineistoon kahdessakin eri muodossa ja vastaavasti eri nimillä.

2. Muunnokset

Kunkin muuttujan skaala voidaan joko pitää alkuperäisenä tai muuntaa joksikin muuksi analyysivaiheessa. Ylemmältä skaalan asteelta voidaan aina mennä alemmalle tasolle. Myös luokittelu- tai järjestysasteikon skaalaa voidaan ja usein tuleekin muuttaa alkuperäisestä. Skaalan muunnos voi tapahtua:

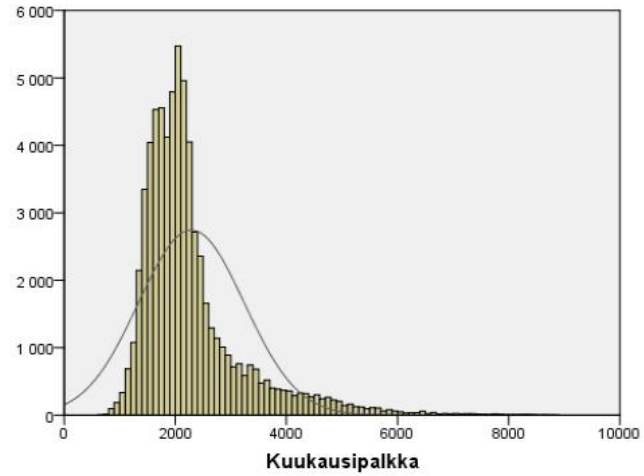
(i) Luokittelulla ja uudelleenluokittelemalla (kategorisointi).
(ii) Lineaarinen muunnos, jolloin alkuperäinen skaala muutetaan toiseksi siten että uudessa muuttujassa säilyvät samat ominaisuudet mutta arvoalue muuttuu. Muuttujan skaala muutetaan sopivalle mukavaksi koetulle alueelle, yleensä samalle kuin jotkin muut saman aihealueen muuttujat. Itse suosin väliä [0, 1] tai [0, 100], jossa 0 on matalin mahdollinen arvo tässä tilanteessa ja 1 tai 100 on vastaavasti korkein mahdollinen.
Summamuuttuja tarkoittaa nimen mukaan muuttujaa joka on ’summaus’ kahdesta tai useammasta alkuperäisestä. Itse teen vastaavan asian summauksen sijasta keskiarvoistamalla yhtenäisesti skaalatuista alkuperäisistä muuttujista. Tällöin siis kaikki muuttujat olisivat samanarvoisia, mutta jos eri muuttujien merkittävyys halutaan erilaiseksi, käytetään keskiarvoa laskettaessa erisuuria painoja. Useammasta muuttujasta painoilla tai ilman muodostettua uutta muuttujaa kutsun yhdistemuuttujaksi.

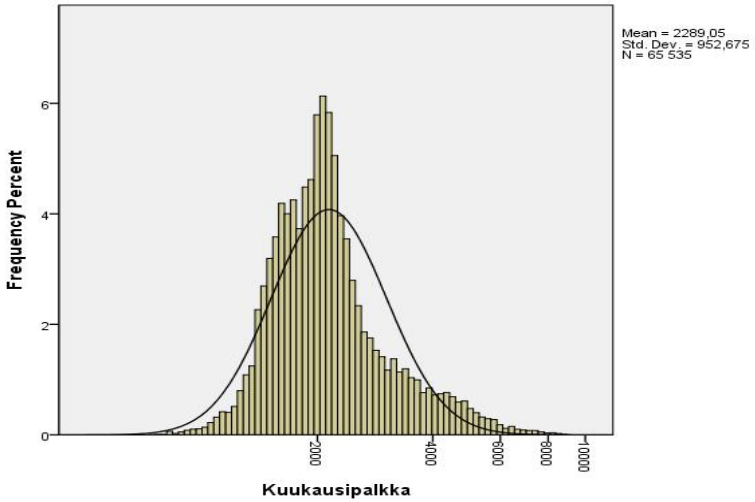
(iii) *Normeeraus tai standardointi:* Jos käytössä on useita suhdeasteikon muuttujia, joita halutaan verrata keskenään ja/tai luoda niistä uusia yhdistemuuttujia, on useita mahdollisuuksia, joista mallittamista ei tässä tarkastella. Muuttujien muunnosmielessä on perusvaihtoehtona *standardointi* siten että arvot pakotetaan lineaarisella (tai muullakin) muunnoksella samalle arvoalueelle, esimerkiksi välille [0,1] tai [0,100], kuten kohdassa (ii) esitin.

Toinen vaihtoehto on *normeeratun normaalijakauman* käyttö eli muodostetaan uusi *z*-muuttuja seuraavasti (s_y on muuttujan *y* keskihajonta).

$$z_k = (y_k - \overline{y}) / s_y$$

(iv) *Logaritminen* muunnos suhdeasteikon muuttujille. Tätä käytetään jos logaritmin ottaminen tulosmuuttujasta tekee jakauman paremmin normaaliseksi, kuten pääsääntöisesti aina tekee jos muuttuja koskee palkkoja, tuloja, saatua toimeentulotukea, tai yrityksen liikevaihtoa ja henkilömäärää, myös indeksilukuja mitä ei aina ymmärretä kuvioissa vaikkapa mediassa ja virallisessa tilastossa.


6 000
5 000
4 000
3 000
2 000
1 000
0
0
2000
4000
6000
8000
10000
Kuukausipalkka

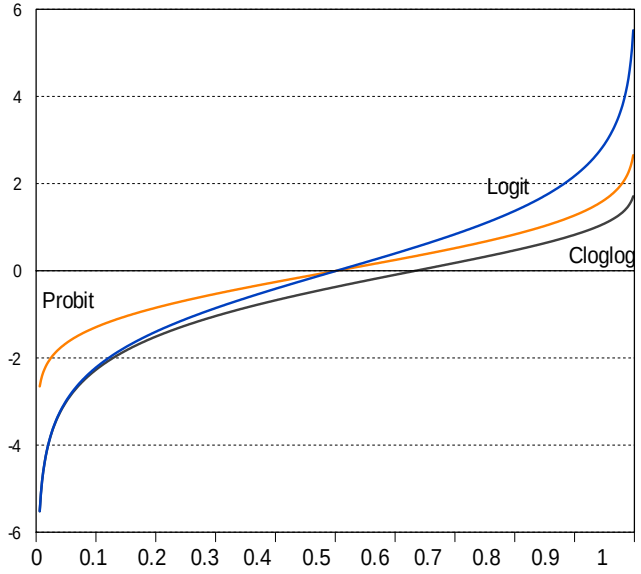

Mean = 2289,05
Std. Dev. = 952,675
N = 65 535
Frequency Percent
6
4
2
0
2000
4000
6000
8000
10000
Kuukausipalkka

(v) *Eksponentiaalinen* muunnos, joka on logaritmisen vastakkaisoperaatio. Se siis voimistaa suurempia lukuarvoja absoluuttisessa mielessä, kun taas logaritminen niitä vaimentaa. Käytännössä sitä harvoin käytetään muutoin kuin palautettaessa logaritmiset arvot alkuperäiselle tasolle. Toki kannattaa kokeilla jos jakauma on sopivasti ylöspäin vino.

(vi) *Logit*-muunnos ja muut todennäköisyysarvojen skaalaukset. Kyseessä ovat välille (0, 1) tai (0%, 100%) sijoittuville muuttujille soveltuvat muunnokset siinä tapauksessa jos väliä ei haluta tarkastella lineaarisena. Tällöin siis kyseessä ovat jo binäärisen selitettävän (saaden arvoja 0 ja 1) muuttujan 'latentit' arvot, jotka sijoittuvat välille (0, 1) tai havaitut tuohon väliin sijoittuvat todennäköisyydet jotka ovat käytännössä lähellä suhteellisia frekvenssejä. Laskulausekkeena muunnos on $log(\pi_k/(1-\pi_k))$ jossa π_k =tarkasteltava todennäköisyys.

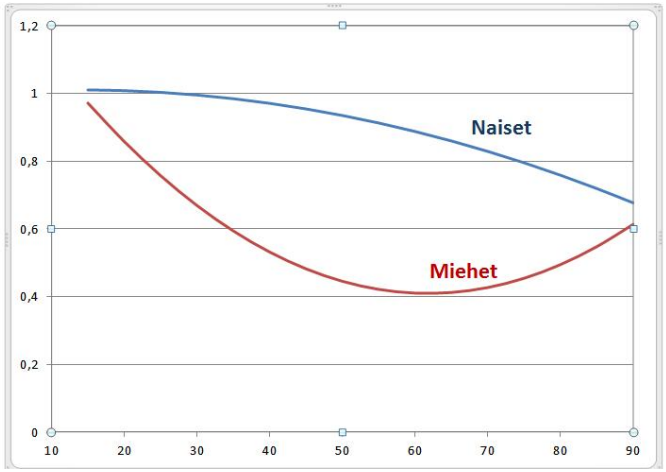
Puhuttaessa eroista tällä skaalalla mitattaessa käytetään sanontoja, 'niin ja niin monta logitia,' ja jos tämä kerrotaan sadalla, puhutaan logit-prosenteista. Ottamalla eksponentti logitista tullaan muotoon $\pi_k/(1-\pi_k)$ jota kutsutaan *vedonlyöntisuhteeksi*.

Logit-skaalasta ei käytännössä paljoa poikkea ns. *Probit*-skaala, joka perustuu kumulatiiviseen normaalijakaumaan. Se on monissa ohjelmistoissa valittavana. Tulkinta on luonnollisesti erilainen, nyt se mittaa todennäköisyyksiä.



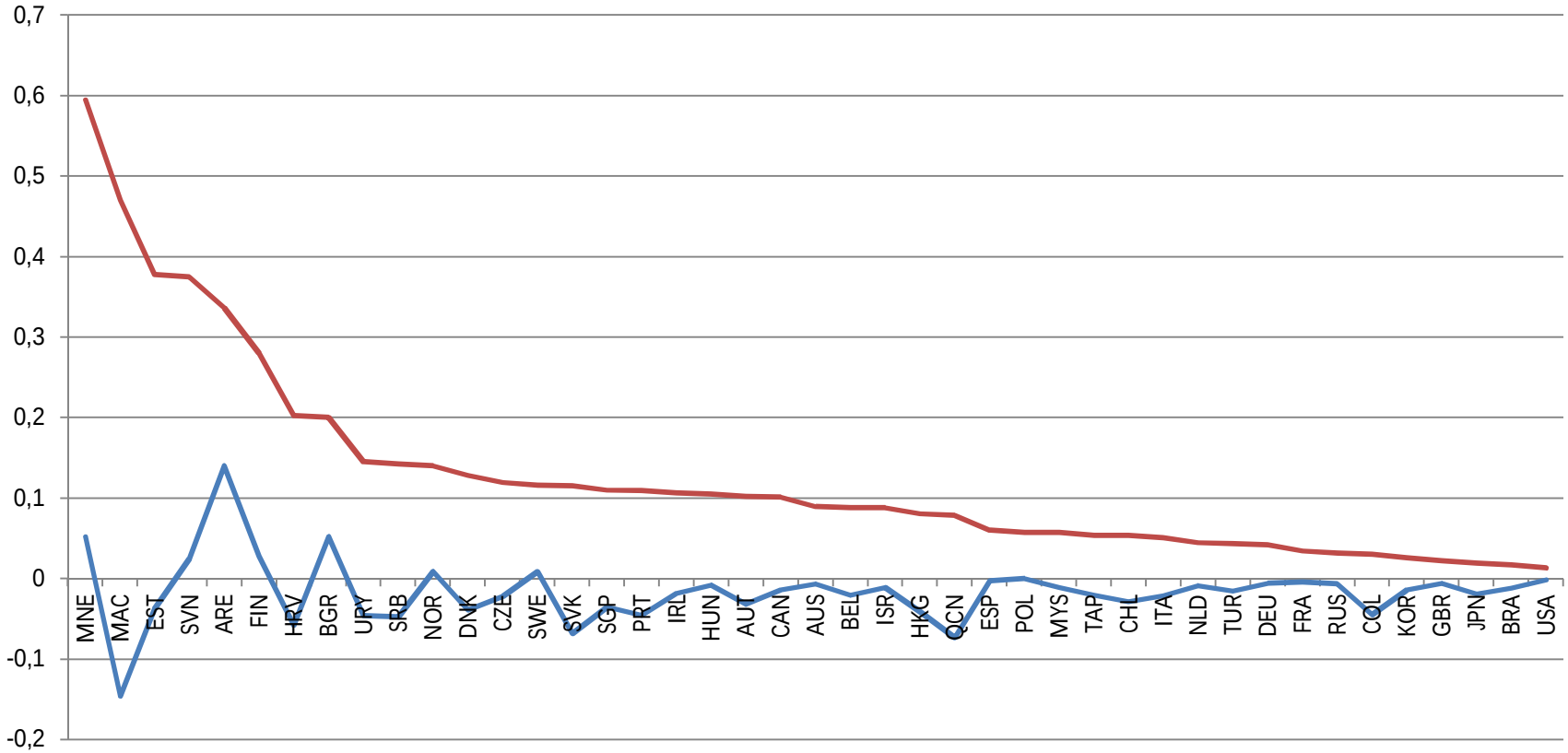
(vii) *Neliöjuuren otto tai neliöön korotus* suhdeasteikon muuttujalle. Vaikutus on samantapainen kuin logaritmoinnilla vs. eksponentoinnilla.

(viii) *Polynominen muunnos*



Esim: Toisen asteen polynomi (siis ikä ja iän neliö mukana) Onnellisuus Suomessa iän mukaan.

Cohenin *d*-esimerkki uudesta Pisasta jossa kaksi osaamista: tekstin ymmärtäminen ja ongelmanratkaisu. Mukana kaikki jälkimmäiseen osallistuneet. Maat järjestetty tekstin ymmärtämistestin eron mukaan (punainen viiva). Maakoodit virallisia. Suomessa tytöt ovat siis paljon parempia, mutta viisi maata on 'edellämme.' Vertaa kirjan tulokseen.



4. Survey-aineiston peruskäsitteistö

Lähden liikkeelle *perusjoukon* käsitteestä. Survey-tutkimuksessa perusjoukko ei ole yksikäsitteinen, vaan tarvitaan viisi käsitettä.

Ensimmäistä kutsun *kiinnostusperusjoukoksi.* Tällainen tutkijaryhmällä on väistämättä mielessä heti hankkeen alussa. Se on ensin aika yleinen mutta pikku hiljaa täsmentyy.

Kun täsmentyminen on edennyt riittävän pitkälle, ollaan valmiita määrittelemään kaikkien empiiristen 'havaintojenkeruututkimusten' avainkäsite eli se joukko, jota todella yritetään tutkia vaikkei ole takeita että sen avulla saadaan kiinnostusperusjoukko hyvin haltuun. Tällaista kutsutaan *tavoiteperusjoukoksi* tai myös ideaali- tai ihanne- tai kohdeperusjoukoksi. On huomattava, että tämän ei tule olla niin ideaali tavoitejoukko, että sitä ei mitenkään voida kunnolla tavoittaa. Siis sen tulee olla silti *realistinen* ja jo mahdollisimman tarkasti rajattu ja aikaan sidottu. Kun tavoiteperusjoukko on valittu ja määritelty, kiinnostusperusjoukko voidaan joksikin aikaa unohtaa.

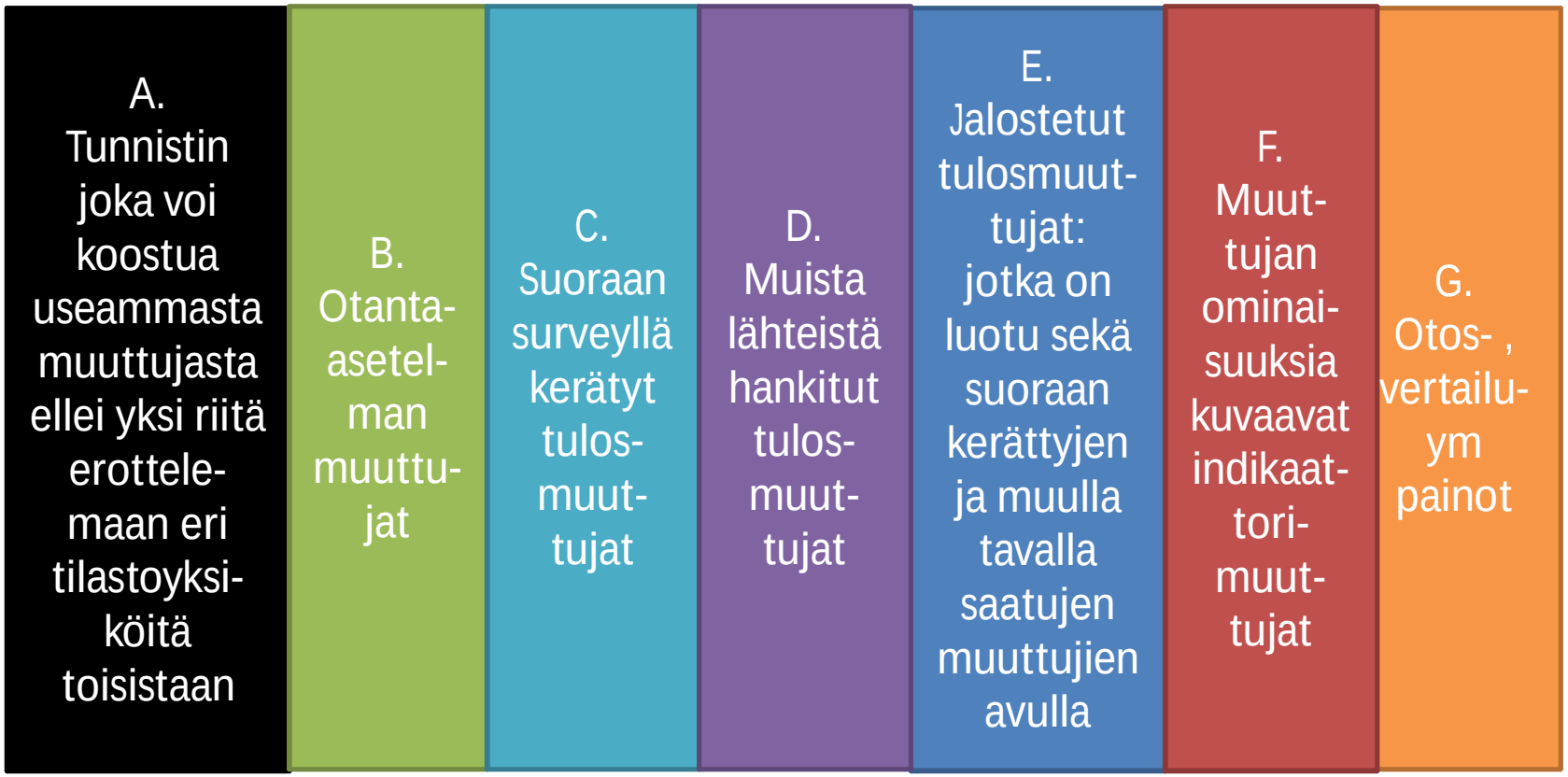
Tavoiteperusjoukkoa yritetään lähestyä käyttämällä sopivaa kehikkoa eli etsimällä paras mahdollinen *kehikkoperusjoukko*, josta poimitaan kaikki tai osa haluttua tiedustelua varten. Kehikkoperusjoukon yksikkö voi olla esimerkiksi henkilö, yritys, kunta, kotitalous, eläin, aika tai alue. Se ei tavallisesti ole saatavissa samalta ajankohdalta kuin miltä itse tiedustelu halutaan tehdä (paitsi jos on aika kehikkona), vaan enemmän tai vähemmän aikaisemmalta ajalta.
Sen vuoksi kehikko usein muuttuu ja tarvitaan kolmas perusjoukko eli *päivitetty kehikkoperusjoukko,* jota käytetään estimoinnissa hyväksi.
Lopulta keräämme itse aineiston, jonka pohjalta me voimme muodostaa ns. *tutkimusperusjoukon*.

Tavoiteperusjoukon ja tutkimusperusjoukon yksiköt ovat ainakin joltakin osin samoja, kutsun niitä *tutkimusyksiköiksi*. Tutkimusyksiköitä voi olla useanlaisia samassa surveyssä, kuten esimerkiksi
- kotitaloussurveyssä kotitaloudet ja kotitalouden jäsenet
tai
- yrityssurveyssä yritykset ja yritysten työntekijät.

Kehikon tai päivitetyn kehikon yksiköt voivat sen sijaan erota olennaisestikin tavoiteperusjoukon yksiköistä. Kehikosta voimme erottaa kahdenlaisia yksiköitä, toisaalta *poimintayksiköitä* joita käytetään tiedustelun otoksen poiminnassa, ja toisaalta *keruuyksiköitä* (joiden sisältä voi vielä löytyä eri raportoijat eli *raportointiyksiköt*) joilta tiedot kerätään. Keruuyksiköitäkin voi olla useanlaisia samassa tiedustelussa. Poimintayksiköiden ja tutkimusyksiköiden erot havainnollistuvat paremmin otannan yhteydessä.

Kaikkien viiden perusjoukon ollessa samat tilanne on helppo. Käytännössä näin ei ole vaan löytyy virhetekijöitä, joista
- kehikkovirheitä ovat *alipeittävyys* (alipeitto), *ylipeittävyys (ylipeitto)*, ja *luokitteluvirheet* kehikkojen osalta; myös *kehikon aito muutos* voidaan eritellä tätä kautta
- vastauskatoa löytyy toisaalta *yksikkö- ja toisaalta erä- eli muuttujatasolta* (*yksikkövastauskato ja erävastauskato*).

Surveyn käyttövalmiin poikkileikkausaineiston yleinen hahmo. Tässä tilastoyksiköiden määrä on sama kuin vastaajien määrä. Surveyn luojan on määriteltävä mikä yksikkö katsotaan vastaajaksi. Merkitsen tätä määrää symbolilla *r* (respondents).



Osa näistä voi olla myös taustamuuttujia, jotka ovat kysyttyjä tai otettu erikseen rekisteristä (esim. ikä, sukupuoli, asuinalue, koulutus, ammatti, toimiala).

Kaavioiden käsitteiden lisätarkasteluja

Yksilöintitunnukset:
Perustunnus (henkilötunnus, yritystunnus, organisaatiotunnus, kuntatunnus, maatunnus, ESS:n kierrostunnus) tarvitaan heti alussa. Sitä käytetään tiedonkeruussa ja pidetään tallessa tiedonkeruuyksikössä. Yleensä tieto on herkkä eikä sitä luovuteta kenelle tahansa. Sen vuoksi voidaan muodostaa uusi, *tietosuojattu tunnus*. Yleinen tapa sen muodostamiseksi on asettaa yksiköt satunnaiseen järjestykseen ja antaa tunnus sen mukaan. Uhritutkimuksessa tehtiin kolme osatunnusta, ensimmäinen alkoi luvusta 10000 ja tuli käyntihaastattelun otokselle, toinen sarja alkoi 20000:sta ja tuli puhelintiedustelun otokselle, kolmas vastaavasti nettiotokselle alkaen luvusta 30000.

Otanta-asetelman ja muut apumuuttujat

Olen ensimmäiseen kaavioon asettanut kolmenlaisiakin apumuuttujia. Ensimmäinen ryhmä B sisältää ne muuttujat joita otanta-asetelmassa käytetään. Näitä on sitä niukemmin, mitä yksinkertaisempi asetelma on. Palaamme näihin otantaa koskevissa luvuissa. On hyvä huomata kuitenkin että koska otos poimitaan kehikosta, joka on lähellä tavoiteperusjoukkoa, nämä muuttujat koskevat laajaa joukkoa, eivät siis otosta. Jälkimmäisessä kaaviossa on otanta-asetelmamuuttujien lisäksi kaksi muuta apumuuttujaryhmää, jotka saatetaan 'unohtaa' lopullisessa vaiheessa koska ne on jo käytetty hyväksi aineiston laadun tarkastelussa sekä hyvien otospainojen muodostamiseksi. Jotkut näistä muuttujista voidaan kuitenkin katsoa myöhemmin tulosmuuttujiksi, jolloin ne ovat ryhmän D muuttujia.

Apumuuttujia kutsutaan myös *X*-muuttujiksi ja lisämuuttujiksi. Niitä on hyvä pyrkiä keräämään mahdollisimman paljon heti surveyn alussa, koska myöhemmin sama työ tulee paljon hankalammaksi ja myös kalliimmaksi, koska voi olla vaikea palata monen kuukauden jälkeen samaan aineistoon takaisin. Apumuuttujia on siis sellaisia jotka ovat perusjoukkotasolta eli sitä koskevia tilastolukuja (väestömäärä sukupuolen, ikäryhmän ja koulutustason mukaan) tai mikrotasoisia eli tilastoyksikön koodeja tai arvoja (sukupuolikoodi, koulutuskoodi, ikä, rekisterin liikevaihto).

Tulosmuuttujat

Surveyaineiston varsinaisia kiinnostuksen kohteena olevia muuttujia kutsutaan *Y-muuttujiksi*, tulos- tai tutkimusmuuttujiksi. Näitä on kahta lajia: ne jotka ovat luonteeltaan samanlaisia kuin alun perin kerätty. Mukana toki voi olla myös sellaisia muuttujia jotka on kerätty aikaisemmin, kuten pitkittäistutkimuksissa, tai muuttujia jotka ovat myös *X-muuttujia.* Kaikista perustason tulosmuuttujista voidaan muodostaa jalostettuja muuttujia joita analyysissä käytetään. Asteikko- ja muunnosliite esittää runsaasti esimerkkejä kummankinlaisia muuttujista. Kannattaa kerrata näitä asioita tässä vaiheessa.

Y-muuttujat ovat pääsääntöisesti siis kerättyjä mutta mukana voi olla myös muuttujien luonnetta kuvaavia indikaattorimuuttujia. Yksi tyyppi näitä ovat ns. *lippumuuttujat* joilla kerrotaan aineiston ja eri muuttujien ominaisuuksista, jolleivät ne muuten ilmene valistuneelle lukijalle. Esimerkiksi, että tällainen muuttuja voi olla kaksiarvoinen kertoen että arvo on imputoitu = 1 tai ei = 0, laatu epävarma= 1 tai ei =0, ja vastaavasti koodaten että on korjattu tai ennustettu.

Meta ja paradata sekä taustadata

Oma erikoislajinsa muuttujia ovat tiedonkeruuseen liittyvät tiedot. Edellä on jo mainittu merkinnät joita haastattelija tekee tiedostoon koskien vastauksen uskottavuutta tai epävarmuutta. Myös haastattelun ajankohta ja kesto on hyvä merkitä aineistoon, kuten myös haastattelijaa itseään koskeva koodi josta pääsee myös haastattelijatietoihin jos on lupa tähän. Luonnollisesti haastattelun tulos myös merkitään ja se kuinka monta kertaa on yritetty vastaajaa tavoittaa ja suostutella tutkimukseen. Tämäntyyppistä tietoa kutsutaan nykyään *paradataksi*. Sillä on suuri merkitys tiedon laadun arvioinnissa ja tulevien tiedustelujen kehittämisessä. Lippumuuttujakin on luonteeltaan paradataa.

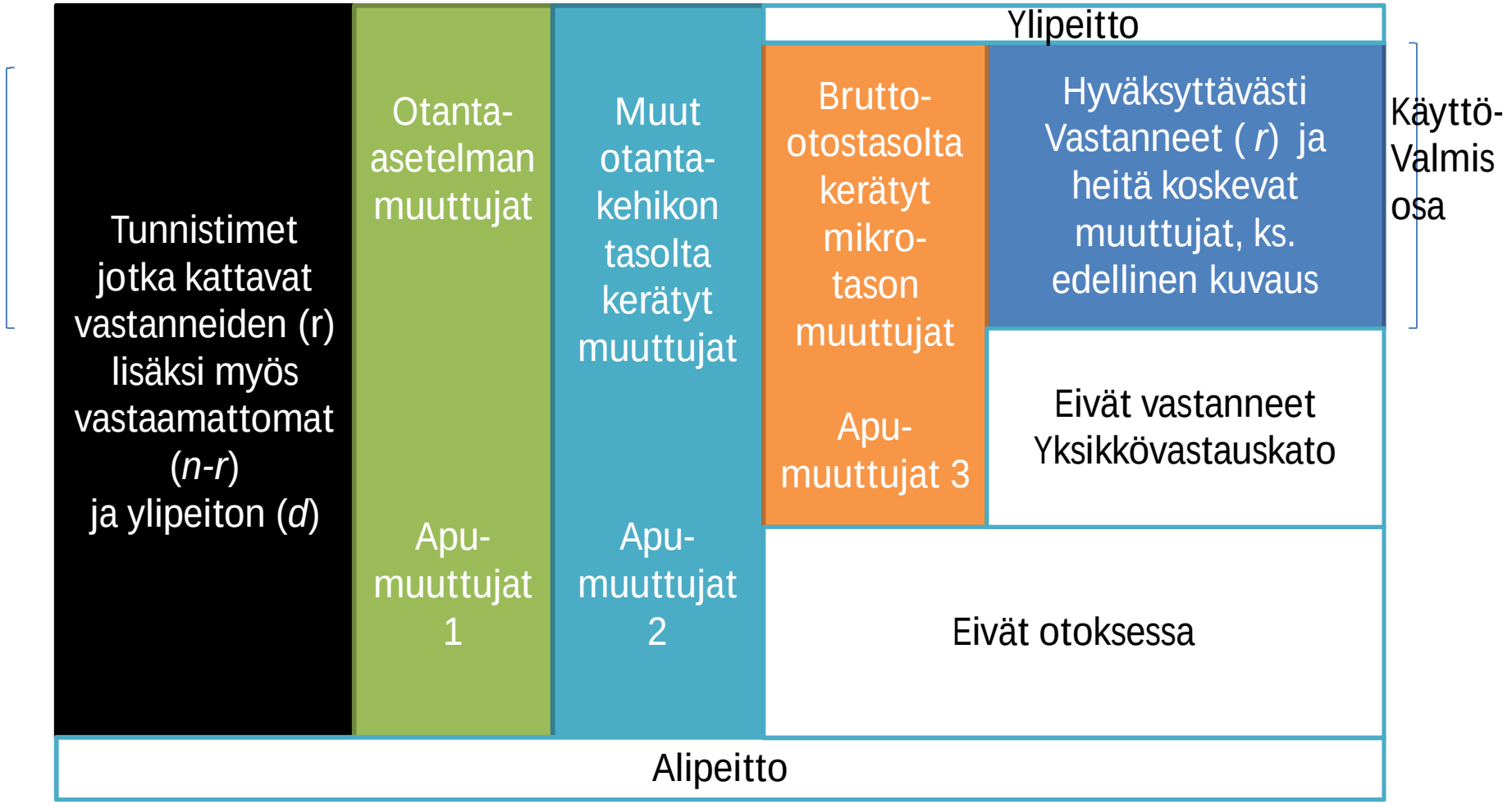
Toinen tietotyyppi on *metadata* joka tarkoittaa tietoa tiedosta. Kyse on siis hyvästä tiedon dokumentoinnista ja mielellään sellaisesta että se voidaan kätevästi hyödyntää aineiston käsittelyssä. Palaan sekä para- että metadataan esimerkeillä useita kertoja. Nykyajan tietotekniikka on jo erinomainen metadatan kuvaamisen kannalta. SPSS koetaan usein helpommaksi kuin SAS mutta kokonaislaatu on molemmissa hyvä ja melkein sama. Luultavasti sitä on SPSS:llä helpompi muodostaa. Sen sijaan *R*-ohjelman huono puoli on sen huono metadatan käyttömahdollisuus.

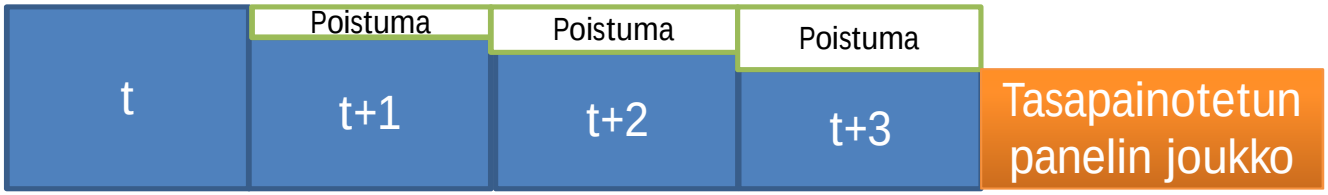
Hyvin toteutetuissa surveyssä on lisäksi *taustadataa*, koskien esimerkiksi surveyn kenttätyöajan tapahtumia ja ilmapiiriä, jopa sääoloja. Tiedustelun keruuaikana tai vähän ennen sitä voi kohdealueella tai maailmassa tapahtua merkittäviä asioita, joilla on vaikutus erityisesti asenteisiin ja myös vastaamishalukkuuteen.

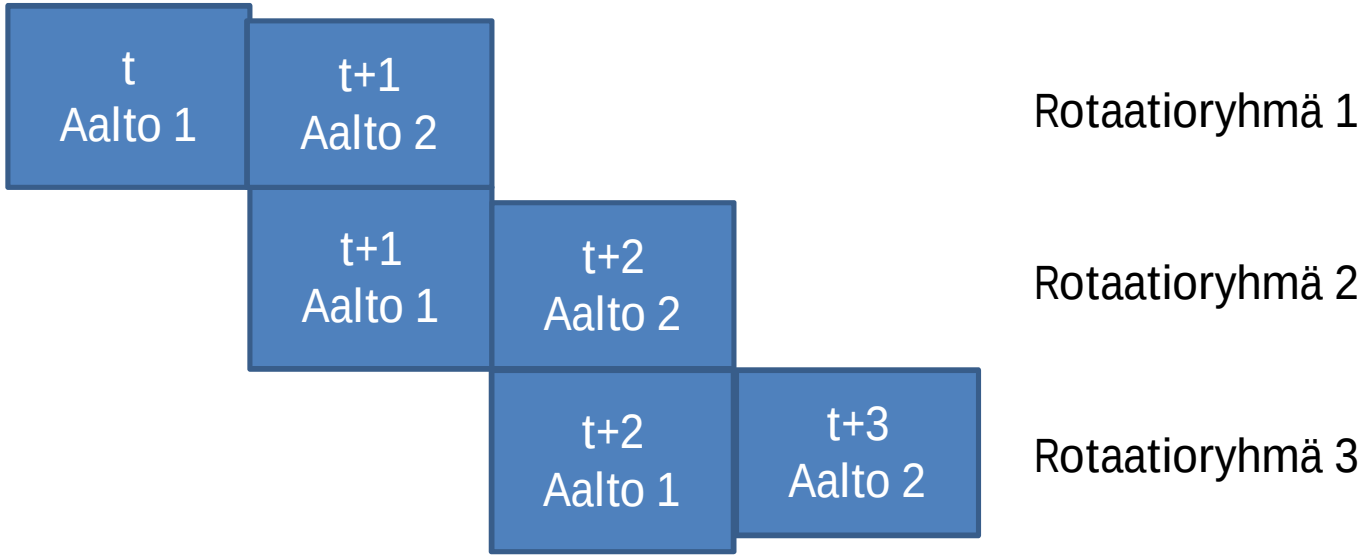
Painomuuttujien osasto
Kuvion loppupäässä on havainnollistettu karkeasti painomuuttujia. Näitä ei tarvita jos koko perusjoukko on mukana aineistossa tai voidaan ajatella että kaikki painot ovat = 1. Kaikissa muissa tilanteissa painot ovat tarpeellisia, myös silloin kun koko perusjoukkoa yritetään tutkia mutta kaikkien tietoja ei saada. Siis jos on vastauskatoa, niin painot ovat tarpeen. Tietysti ne ovat vielä välttämättömämpiä, jos vastaajakandidaatit on valittu otoksella. Painoja voidaan kutsua tällöin otospainoiksi, mutta otospainoja on useanlaisia.

Aluksi luodaan kehikkotietoja käyttäen *asetelmapaino*. Kun aineisto saadaan koottua, on mahdollista luoda ensiksi otanta-asetelman ja saatuun aineistoon perustuva uusi paino, jota kutsun *peruspainoksi* tai *perusotospainoksi*. Jos on käytettävissä muutakin tietoa kuin alkuperäisen kehikon tiedot, niin on mahdollista luoda parempia, ns. *adjustoituja, vastauskato-oikaistuja otospainoja*. Otospainoja käytetään lähes kaikessa analyysissä. Aineiston käsittelijän tehtävänä on valita kuhunkin tilanteeseen sopivin otospaino. Analysoijan ei tarvitse viimeisen päälle ymmärtää miten paino on muodostettu mutta on syytä olla tiukkana ja vaatia hyvät vakuudet sen hyvyydestä tiedon kerääjältä.

Surveyn poikkileikkausaineiston yleinen hahmo käyttövalmiin tiedoston taustalla




Poistuma
Poistuma
Poistuma
t
t+1
t+2
t+3
Tasapainotetun panelin joukko


t
Aalto 1
t+1
Aalto 2
t+1
Aalto 1
t+2
Aalto 2
t+2
Aalto 1
t+3
Aalto 2
Rotaatioryhmä 1
Rotaatioryhmä 2
Rotaatioryhmä 3