Tilastokeskus

Otantamenetelmien lähihistoriaa;
Otantamenetelmän kehitys USA:ssa 1930-1950

Vesa Kuusela

Haastattelu- ja tutkimuspalvelut

Tilastokeskus

Todennäköisyysotanta

- Jokaisella perusjoukon jäsenellä on **tunnettu** todennäköisyys (>0) tulla valituksi otokseen
  - Sisältymistodennäköisyyksien ei tarvitse olla samoja
    - Teoriassa jokaisella perusjoukon jäsenellä voisi olla eri todennäköisyys (Horvitz-Thompson –estimaattorit) $\hat{\bar{y}} = \sum_{i=1}^{n} y_i / Np_i$
- Sisältymistodennäköisyyksien laskenta edellyttää, että perusjoukosta on olemassa hyvin määritelty kehikko
  - Mitä monimutkaisempi otanta-asetelma sitä enemmän tarvitaan tietoa otosyksiköistä
  - Kehikosta on löydyttävä tietoa myös otosyksikön saavuttamiseksi
  - Jos kehikossa ei ole riittävästi tietoja, se voidaan kompensoida monivaiheisella otannalla

Tilastokeskus

Nykyaikaisen otantatekniikan kehittäminen

- Nykyinen otantateoria kehitettiin pääosin Yhdysvalloissa 1930- ja 1940-luvulla käytännön tarpeisiin
  - Merkittävin kehitys tapahtui USA:n tilastovirastossa (Bureau of the Census) 1940-luvulla
  - Merkittävää kehitystä tapahtui Englannissa 1930-luvulla, mutta monet keskeiset henkilöt muuttivat USA:han ennen 2. maailmansotaa
- Syitä kehitykselle USA:ssa oli useita:
  - Yhdysvalloissa on hyvin pitkä otostutkimusten perinne
    - *The Harrisburg Pensylvanian* teki vuonna 1824 ensimmäisen ns. ”straw poll” selvityksen; vuonna 1883 *Boston Globe* kehitti ovensuukyselyn
    - Maatalous-surveyt yleistyivät 1800-luvun loppupuolella
    - Useat sanomalehdet tekivät mielipidetutkimuksia jo 1900-alkupuolella
  - Monissa yliopistoissa koulutettiin tilastotieteilijöitä
  - Iowa State University Statistical Laboratory
  - Suuri Lama

Tilastokeskus

Iowa State University Statistical Laboratory

- Snedegor perusti 1930-luvun puolivälissä Amesiin
- Malli otettiin Rothamstedin vastavasta laboratoriosta, jossa R.A. Fisher oli päätilastotieteilijänä
- Kaikkien empiiristen tutkimusten analysointi oli laboratorion vastuulla (ainoa laitos maailmassa siihen aikaan jolla oli tällainen asema)
- Muiden muassa Cochran, Sukhatme ja Kempthorne siirtyivät Englannista Amesiin
- Keskittyi maatalouden tutkimuksiin otantamenetelmillä (Jessen)
  - Alueotanta ja paneelien rotatointi
- Yksi keskeinen tutkimusala oli tilastollinen tietojenkäsittely
  - Rakensivat ensimmäisen laskentalaitteen, jolla oli tietokoneen keskeiset osat

Tilastokeskus

Yhdysvaltain Suuri Lama ja New Deal -ohjelmat

- Lokakuun 29. päivänä New Yorkin pörssi romahti
  - Osakkeiden arvosti hävisi noin 90%
  - Seurasi Suuri Lama
    - Yhdysvaltain teollisuustuotanto väheni 49% ja BKT 39%
    - Työttömiä oli yli 20% ja joillain alueilla yli 60%
- Presidentti Franklin D. Rooselvelt käynnisti ns. New Deal –ohjelman työttömien olojen helpottamiseksi
  - New Deal vaati tuekseen tietoja väestön tilasta, mutta niitä ei ollut
    - Arviot työttömien määrästä vaihtelivat 5 – 13 milj. välillä
  - Useita virastoja perustettiin ohjelman toteuttamiseksi, mm. Work Project Administration (WPA)

Tilastokeskus

Otostutkimusten tekemisen käytännön ongelmia

- Tietojen keruu edellytti haastattelemista
  - Haastattelijat paikallistivat vastaajat, koska ei ollut sellaisia otoskehikkoja, joista olisi selvinnyt vastaajien nimet
    - Haastattelijoiden käyttö edellyttää haastatteluorganisaatiota, joka on työläs perustaa ja kallis ylläpitää
  - Luku- ja kirjoitustaito ei ollut yleistä
- Ei ollut tietojenkäsittelyä nykyisessä mielessä
  - Reikäkorttikoneet tulivat laajempaan käyttöön 1930-luvulla, mutta niiden käyttö oli hidasta ja työlästä
  - Otosten poiminta piti tehdä manuaalisesti

Tilastokeskus

Muita otossuunnittelun ongelmia 1930-luvulla

- Perusjoukosta oli vain vähän ja päivittämättömiä tietoja
  - Väestölaskenta tehtiin joka 10. vuosi ja tiedot oli paperilomakkeilla
  - Tietoja/talous ei kerätty kovin paljon eikä niitä päivitetty
- Ei hallittu muita otantamenetelmiä kuin YSO (SRS)
  - YSO:n poiminta väestönlaskennan tiedoista olisi ollut työlästä
  - YSO olisi johtanut hitaaseen ja kalliiseen tietojen keräämiseen
- Hallinnossa ei uskottu, että otostutkimukset voisivat tuottaa luotettavaa tietoa
  - Census Bureaun johto pelkäsi, että otostutkimukset pilaisivat viraston maineen
- Jotkin virastot tekivät otospohjaisia tutkimuksia, mutta niiden menetelmiä arvosteltiin voimakkaasti tutkijapiireissä

Tilastokeskus

Roosevelt – Langdon -presidentinvaalien ennusteet

- *The Literary Digest* lähetti 10 miljoonaa ”äänestyslippua”
  - Näistä palautettiin 2 miljoonaa
  - Ennuste: Alfred Langdon voittaa 57% vs. 43%
- George Gallupin perustama *American Institute of Public Opinion* …
  - Käytti koulutettuja haastattelijoita ja juuri kehitettyä kiintiöpoimintaa
  - Teki 3000 haastattelua ympäri maata
  - Ennuste: Franklin Roosevelt voittaa 54% vs. 46%
- Vaalissa Franklin D. Roosevelt voitti 61% ääniosuudella
- *The Literary Digestin* ”kutsu” lähetettiin lehden omille lukijoille, auton omistajille ja puhelimen omistajille
  - Lisäksi vastaaminen edellytti omaa aktiivisuutta, mikä lisäsi valintaharhaa

Tilastokeskus

Tutkimuksia työttömyyden tutkimukseksi

- Vuonna 1937 tehtiin WPA:n aloitteesta vapaaehtoisuuteen pohjautuva työttömien kirjaaminen.
  - Jokaiseen kotitalouteen jaettiin (Rooseveltin allekirjoittama) lomake, jossa pyydettiin rekisteröitymään, jos oli työtön.
  - Kenttätyön toteutti kansallinen postilaitos.
- Tulosten luotettavuuden arvioimiseksi tehtiin tarkistuslaskenta (*enumerative check census*) satunnaisotantaa käyttäen
  - Ensiksi poimittiin satunnaisesti joukko maantieteellisiä alueita ja näistä haastateltiin jokainen kotitalous postin jakelureiteillä
  - Postinkantajat toimivat aluksi haastattelijoina ja sitten yhdistivät haastattelut ja aikaisemmat vapaaehtoiset ilmoitukset
  - Tämän tutkimuksen onnistuminen vakuutti maan hallinnon otostutkimusten käyttökelpoisuudesta.

Tilastokeskus

Työttömyyden tutkimus siirtyy Census Bureaulle

- Vuona 1940 WPA alkoi suunnitella kuukausittain tehtävää työttömyyttä mittaavaa otostutkimusta (*Sample Survey of Unemployment*)
  - Otanta-asetelmana oli monivaiheinen otanta
  - Menetelmä oli satunnainen muuten, mutta viimeisen vaiheen yksiköt valittiin maantieteellisesti siten, että se ei johtanut talouksien täysin satunnaiseen valintaan
- Vuonna 1942 tutkimus siirrettiin Census Bureaun tehtäväksi ja sen nimeksi muutettiin **Current Population Survey** (CPS)
- Menetelmää alkoi kehittää uusi juuri valmistunut tutkijapolvi

Tilastokeskus

Jerzy Neymanin panos

- Neyman esitti RSS:n kokouksessa vuonna 1934 teorian, johon nykyinen otantateoria nojaa (Neyman 1934, JRSS,97,558-606)
  - Perusjoukon parametrit oletetaan vakioiksi (ei tarvita á priori jakaumaa)
  - Otosjakauma muodostuu (oletetuista) toistetuista otoksista samasta perusjoukosta
  - BLUE (Best Linear Unbiased Estimator)
  - Väliestimaatit, luottamusvälit
- Neymanin vuonna 1937 julkaisema artikkeli tarkensi teoriaa eiryisesti todennäköisyyslaskennan osalta (Neyman 1937, Phil. Trans. Royal Soc., 333-380)
- Vuonna 1938 Neyman julkaisi artikkelin, jossa hän esitteli kaksivaiheisen otannan ja optimaalisen osituksen (Neyman 1938, JASA, 33, 101-116)
  - Ensimmäistä kertaa puhuttiin eri suuruisista sisältymis-todennäköisyyksistä
  - Menetelmä oli tarkoitettu otannan tekemiseen ison ja moni-ilmeisen maan väestöstä, kun tietojen kerääminen edellytti pitkää haastattelua

Tilastokeskus

Sampling from human population

- Lähes kaikki otannan tutkimukset 1930-luvulla käsittelivät maatalouden tutkimuksia
- Neyman vieraili USA:ssa 1937 ja hänelle esitettiin kysymys väestön tutkimisesta, johon hän ei osannut suoralta kädeltä vastata.
  - Tutkittavan asian, *x*, selville saaminen vaatii pitkän (ja kalliin) haastattelun
  - Kustannusten on pysyttävä hyväksyttävinä
  - Ilmiön *x* ja *y* välillä on voimakas korrelaatio ja *y* voidaan selvittää pienin kustannuksin
  - Ensi vaiheessa selvitetään *y*:n jakauma ja perusjoukko ositetaan sen mukaan
  - Kustakin ositteesta poimitaan pieni otos *x*:n selvittämiseksi

Tilastokeskus

Maataloustutkimus Iowassa

- Raymond Jessen selvitti otantamenetelmän soveltuvuutta ja siihen kytkeytyviä virheitä maataloustuotannon selvittämisessä
  - Kaksi samanlaista surveytä,1938 ja 1939
  - Otos poimitaan satunnaisesti kartan (ilmavalokuva) päälle piirretystä ristikosta. Yksi alue kustakin kunnasta.
  - Laskijat haastattelivat jokaisen farmarin alueelta
  - Vuoden 1939 tutkimuksessa haastateltiin 50% edellisellä kerralla mukana olleista maanviljelijöistä
- Tutkimusta käytettiin mallina myöhemmin aluepoiminnan ja rotatoivien paneelien kehittämisessä

Tilastokeskus

Otanta-asetelma Current Population Surveylle (CPS)

- Census Bureaun tilastotieteilijät M. Hansen, W. Hurwitz, W. Madow saivat 1940-luvun alussa tehtäväksi kehittää toteuttamiskelpoisen otantasuunnitelman CPS:lle
- Lähtökohdaksi otettiin Neymanin 1938 artikkeli ja Jessenin maataloustutkimuksen asetelma.
- Tulos julkaistiin 1943: Hansen, M. H., & Hurwitz, W. N. "On the theory of sampling from a finite population," *Annals of Mathematical Statistics*, 14, 333-362.
- Hansenin ja Hurwitzin menetelmä oli kaksivaiheinen PPS-otanta palauttaen. Vuonna 1952 Horvitz ja Thompson laajensivat sitä otantaan palauttamatta
- Menetelmän edut: tiedon keruun kustannukset olivat hyväksyttävät, tiedonkeruun kenttätyön organisointi olis helppoa, haastattelijoiden työkuorma pysyi tasaisena, estimaatit olivat riittävän tarkkoja

Tilastokeskus

CPS:n otanta-asetelma nykyisin

- Monivaiheinen ositettu otanta
  - 56 000 asuntoa (osoitetta) 792 otanta-alueelta poimitaan edellisen väestönlaskennan pohjalta
- Ensimmäinen vaihe
  - Maa jaetaan maantieteellisesti ensivaiheen otosyksiköihin (PSU)
  - PSUt jaotellaan taloudellisesti ja yhteiskunnallisesti homogeenisiin osittеisiin
  - Kustakin osittеesta yksi PSU poimitaan PPS-otannalla (suhteessa osittеen asukasmäärään)
- Toinen vaihe
  - Poimitaan otos asunnoista (osoitteista) ensi vaiheessa poimituista PSU:sta
  - Lopulliset otosyksiköt (USU) ovat n. neljän asunnon ryppäitä
  - Otos poimitaan systemaattisesti valittujen USU-alueiden osoitelistata
- Jokainen talous haastatellaan kuukausittain seuraavan 4 kuukauden aikana, sitten 8 kuukauden taukoja sitten taas 4 kertaa kuukauden välein

Tilastokeskus

Nykyaikaisen otantateorian ensimmäiset oppikirjat

- Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1953), *Survey sampling methods and theory, Vols. I and II*: John Wiley & Sons
- Cochran, W. (1953), *Sampling Techniques*: John Wiley & Sons.