Tilastollinen tutkimus ja raportointi

Sisältää otteita Esa Läärän artikkelista ”Tilastollisen tutkimuksen Raportointi”, 2005 (kirjoittajan luvalla).

Motto*: “Statistics is about common sense and good design”* (Campbell ja Machin: *Medical Statistics - A Commonsense Approach*. Wiley 1993)

Raportin rakenne

1. Tutkimusongelman ja -asetelman määrittely.
2. Populaation kuvaus.
3. Metodologia.
4. Keskeiset tulokset.
5. Tutkimusongelmien ja tavoitteiden arviointi.
6. Kritiikki, populaation saavuttaminen, kato.

Tutkimusongelman ja –asetelman määrittely, hypoteesit

- Mitä? Millaisia ovat perheiden kulutustottumukset?
- Mitä eroa? Mitä eroa on kulutustottumuksilla eri ikäluokissa?
- Miten ilmenee? Miten huono yrityskuva ilmenee käytännössä?
- Miten vaikuttaa? Miten vuodenajan vaihtelut vaikuttavat urheiluvälineiden myyntiin?

Tutkimushypoteesi on em. tekijöihin liittyvä väite.

Tutkimusasetelma on joko kokeellinen tai havainnoiva.

Tutkimuspopulaation kuvaus

- Joukko ihmisiä, yrityksiä, tapahtumia,...
- Populaation rajaus tärkeä: Esim.

  ”Tuotetta X käyttävät henkilöt”

  ”Viime vuonna tapahtuneet konkurssit”

  ”Turkulaiset palveluyritykset”.

Havainto- ja mittausmenetelmien kuvaus

- Tietojen keruumenetelmät: www-kyselyt, henkilökohtaiset haastattelut, postikyselyt, puhelinhaastattelut, valmiit tietokannat,...
- Otoksen valintamenetelmät: satunnaiset ja ei-satunnaiset otantamenetelmät.

Tilastollisen kuvailevan analyysin menetelmät

- Taulukot.
- Kuviot.
- Keskiluvut, hajontaluvut, muut.
- Riippuvuusanalyysi.

Tulokset: Kategoriset muuttujat

- Luokitteluasteikollisen muuttujan (*nominal scale*) jakauman kuvaamisessa käytetyin tunnusluku on kuhunkin luokkaan sijoittuvien havaintoyksiköiden suhteellinen osuus (*relative frequency*, *proportion*).
- Kaksiluokkaisen (*dichotomous, binary*) muuttujan tapauksessa (esim. 1 = ”yritys menee konkurssiin” *vs*. 0 = ”ei mene”) tarvitaan vain toisen luokan suhteellinen osuus.
- Graafisesti luokittelu-tason muuttujan jakaumaa esitetään usein mm. pylväs- (*bar chart*) ja piirakkakuvioilla (*pie chart*), joista erityisesti jälkimmäistä kehotetaan tieteellisissä esityksissä yleisesti välttämään.
- Myös järjestysasteikollisen muuttujan (*ordinal scale*) jakaumaa voidaan kuvata luokkakohtaisilla suhteellisilla osuuksilla ja em. piirroksin mutta lisäksi myös kumulatiivisin osuuksin.

Taulukot

- (T1) Taulukon on oltava mahdollisimman yksinkertainen ja itsensä selittävä. Se on kyettävä ymmärtämään "Aineisto ja menetelmät"-lukuun perehtymisen jälkeen ilman tukeutumista tekstiin.
- (T2) Taulukossa on otsikko, joka kertoo, mistä aineistosta tai sen osasta on kysymys ja mitä tietoja taulukko sisältää.
- (T3) Lähde on mainittava, jos tiedot ovat peräisin muusta julkaisusta tai aineistosta, jota artikkelissa ei pääsääntöisesti käsitellä.
- (T4) Jokaisella rivillä ja sarakkeella on tiivis yksikäsitteinen nimi tai otsikko (esim. ikävuosien luokitus 0-15, 15-25, 25-35, ... ei ole yksikäsitteinen mutta 0-14, 15-24, 25-34, ... on).
- (T5) Käytetyt mittayksiköt, koodit ja lyhenteet on selvitettävä, tarvittaessa alareunaan sijoitettavissa viitteissä.
- (T6) Sarakkeiden otsikot ja alaviitteet erotetaan muusta taulukosta vaakasuorin viivoin. Eri hierarkiatasojen sarakeotsikot erotetaan myös toisistaan vaakaviivoin siten, että ylemmän tason otsikon alle tuleva yhtäjaksoinen viiva kattaa vain tälle otsikolle alisteiset sarakkeet.
- (T7) Sarakkeita ei nykyisin vallitsevan käytännön mukaan erotella pystyviivoilla. Tässä tosin julkaisusarjojen käytännöt hieman vaihtelevat.

Taulukot

- (T8) Vertailtavat suureet kannattaa sijoittaa mahdollisimman lähelle toisiaan. Lukujen vertailu allekkain on helpompaa kuin rinnakkain.
- (T9) Tyhjiä soluja ei jätetä. Jos tieto puuttuu, se korvataan kahdella pisteellä. Jos lukua ei ole mielekkäästi olemassa merkitään yksi piste. Yhdysviiva tarkoittaa, että ao. lukumäärä tai siihen perustuva suhteellinen osuus on nolla, mutta luku 0 kuvaa arvoa, jonka suuruus on alle puolet pyöristystarkkuudesta (esim. alle 0.5 %).
- (T10) Taulukoitaessa luokiteltuja muuttujia ristiin prosenttilaskujen suunnan yleisperiaate on se, että esitetään selitettävän muuttujan arvojen prosenttijakaumat selittävän tai ennustavan muuttujan luokissa, jos otanta-asetelma sallii. Prosenttien nimittäjien täytyy myös näkyä esim. omilla riveillään tai sarakkeillaan (usein suluissa).
- (T11) Erityisesti lukumäärä- ja prosenttijakaumien yhteydessä on hyvä laskea näkyviin myös summarivejä ja -sarakkeita. Periaatteessa kaikkien yhteenlaskettavien tulisi olla näkyvissä, eli myös luokan "tieto puuttuu" (joskaan tälle luokalle ei aina välttämättä tarvitse varata omaa riviä tai saraketta, kunhan ainakin alaviitteessä ilmaistaan, kuinka moni puuttuvan tiedon omaava havaintoyksikkö sisältyi ao. rivi- tai sarakesummaan).
- (T12) Taulukon luettavuutta parantaa usein, jos rivit ja sarakkeet järjestetään niiden "suuruuden" mukaisesti (esim. reunajakaumien, summarivien tai sarakkeiden lukujen perusteella, suurinta lukua vastaava luokka ylimmälle riville tai vasemmanpuolimmaiseksi sarakkeeksi, jne.), ellei luokkien järjestystä täysin määrää painavampi sisällöllinen kriteeri.

Kuviot

- Em. ohjeet (T1) (T3) ovat suoraan käännettävissä vastaaviksi kuvaohjeiksi.
- (K4) Kuvan tyypin valinnassa ja sen yksityiskohtien toteutuksessa kannattaa tähdätä ns. data/muste-suhteen (*data/ink ratio*) maksimointiin eli käyttää viivaa, symboleita, varjostuksia ym. säästeliäästi olennaisen informaation välittämiseksi.
- (K5) Kuva-alue täytetään mahdollisimman tehokkaasti. Akselin asteikko ulotetaan hieman ao. muuttujan havaitun vaihteluvälin ulkopuolelle. Tehokas täyttäminen voi joskus vaatia asteikon katkaisua, jota pitää kuitenkin käyttää varoen.
- (K6) Vaaka- ja pystyakselien kuvaamat suureet mittayksiköineen on nimettävä selvästi. Asteikon jakopisteitä merkitään riittävästi (minimi on kolme) mutta ei liian tiheästi. Havaitun vaihteluvälin ulkopuolisia jakopisteitä on yksi kummallakin puolella. Jakoviivat merkitään kuva-alueesta ulospäin.

Kuviot

- (K7) Lukumääriä, suhteellisia osuuksia ym. ei-negatiivisia suureita kuvaava asteikko alkaa periaatteessa nollasta (ks. kuitenkin (K5)).
- (K8) Kuva-alue kannattaa usein kehystää kopioimalla vaaka-akseli alueen yläreunaan ja pystyakseli oikeaan reunaan, jakoviivat edelleen ulospäin. Jomman kumman akselin muutamia tärkeitä "viitearvoja" voi paikallistaa toisen akselin suuntaisilla ohuilla viivoilla yli koko kuva-alueen.
- (K9) Käytettävät symbolit, viivatyypit, rasterit, varjostukset, koodit, lyhenteet, tunnukset ym. on selitettävä lyhyesti joko otsikon yhteydessä tai tiiviillä avaimella, joka on selvästi erillään datan täyttämän alueesta.
- (K10) Eroteltaessa eri luokkia, ryhmiä ym. toisistaan täytyy käyttää hyvin toisistaan erottuvia symboleita, viivoja ym.
- (K11) Erityisen kriittinen pitää olla kolmiulotteisen vaikutelman antavien kuvien suhteen; niitä ei varsinkaan pidä käyttää silloin, kun kolmas ulottuvuus ei pidä sisällään erillistä sisällöllistä ulottuvuutta eli uutta muuttujaa.

Tulokset: Numeeriset muuttujat

- Jatkuvan (*continuous*) välimatka- (*interval scale*) tai suhdeasteikkotasoisen (*ratio scale*) muuttujan jakauman kuvaamiseen on mahdollista käyttää luokiteltua prosenttijakaumaa taulukoituna kuten edellä.
- Se voidaan graafisesti esittää isolla aineistolla histogrammina (*histogram*, eli pylväskuvio, jossa vierekkäisten luokkien pylväät ovat toisissaan kiinni ja pylvään pinta-ala on suoraan verrannollinen ao. luokan suhteelliseen osuuteen eripituisia luokkia käytettäessä).
- Kun havaintoja per ryhmä on vähän (esim. alle 20), havainnollisempi esitys saadaan pistekuviolla (*stripchart* tai *dotplot*) ja keskisuuren ryhmäkoon tapauksessa myös laatikko-janakuviolla (*boxplot*).

Sijainnin ja hajonnan tunnusluvut

- Jatkuvan muuttujan jakaumaa on myös tapana kuvata tiiviisti sen sijainnin ja hajonnan tunnusluvuilla. Sijaintiluvuista (*measures of location*) käytetyimmät ovat aritmeettinen keskiarvo (*mean*) ja mediaani (*median*). Muuttujan arvojen vaihtelua kuvaavista hajontaluvuista (*measures of dispersion*) tavallisimpia ovat keskihajonta (*standard deviation*, lyhennetään usein s tai SD), vaihteluväli (*range*) eli pienin ja suurin arvo sekä kvartiiliväli (*interquartile range*) eli ala- (*lower quartile*) ja yläkvartiilin (*upper quartile*) muodostama lukuväli. Mediaani, kvartiilit ja äärimmäiset arvot ovat myös laatikkokuvion perusainekset.

Riippuvuus ja korrelaatio

- Kahden vähintään välimatka-asteikollisen muuttujan välistä tilastollista riippuvuutta näkee usein kuvattavan Pearsonin korrelaatiokertoimella (*correlation coefficient*) *r*. *r* on usein epäinformatiivinen ja jopa pahoin harhaanjohtava. Sen arvo riippuu suuresti kummankin muuttujan jakauman hajonnasta ja vinoudesta, niiden yhteisjakauman ja riippuvuuden muodosta sekä yksittäisistä ns. vieraista (*outlier*) tai vaikuttavista (*influental*) havainnoista, joita koskevat lisätiedot ovat tarpeen *r*:n arvoa tulkittaessa. Monet näistä piirteistä välittyvät hyvin sirontakuvion (*scatter plot*) avulla, jonka esittäminen on suotavaa aina kun kiinnostavia jatkuvien muuttujien riippuvuuksia halutaan havainnollistaa.

Aikasarjat

- Toistettujen mittausten (*repeated measurements*) tai mittaussarjojen (*serial measurements*) asetelmassa samaa muuttujaa mitataan toistuvasti tietyn aikataulun mukaan (esim. toistetaan kysely vuoden välein).
- Tällaista tutkimusta nimitetään myös pitkittäistutkimukseksi (vrt. poikittaistutkimus).

Satunnaisvirheen arviointi — tilastollinen päättely

- Tilastollisen päättelyn avulla arvioidaan tulosten yleistettävyyttä.
- Hypoteesin testaus: Onko muuttujien välinen korrelaatio otoksessa riittävä merkki siitä, että sitä on myös populaatiossa?
- Luottamusvälit (virhemarginaalit): Kuinka paljon miesten ja naisten keskipalkoissa on populaatiossa eroa?

Kritiikki

- Varsinkin kyselytutkimuksissa katoanalyysi on tärkeää.
- Katoanalyysissä pyritään selvittämään, onko kyselyyn vastanneiden ja ei-vastanneiden joukossa merkittävää eroa, ja onko erolla mahdollisesti merkitystä tulosten kannalta.
- Kausaliteettien arviointi.

Monimuuttujamenetelmiä

- Havaintojen ryhmittely: Klusterianalyysi
- Muuttujien ryhmittely: Faktorianalyysi
- Keskiarvojen vertailu: Varianssianalyysi
- Numeerisen muuttujan selittäminen: Regressioanalyysi
- Kategorisen muuttujan selittäminen: Logistinen regressioanalyysi.

Lähteitä tutkimussuunnitteluun

- Anttila, P. & Kataikko, M.-S. & Tenkama, P. (toim.) 2000. Tutkimisen taito ja tiedon hankinta. Hamina. Akatiimi Oy.
- Hirsijärvi, S. & Remes, P. & Sajavaara, P. 2004. Tutki ja kirjoita. Jyväskylä. Gummerus Kirjapaino Oy.