

Huomaa että käsitteet Tilastollinen Merkitsevyys ja Yhteiskunnallinen tai muu **Merkittävyys** ovat eri käsitteitä.

Edellinen kertoo riittävän vankasta estimaattien erosta tietyssä tavoiteperusjoukossa kun käytettävissä on jokin osa-aineisto, surveyssä otos. Tilannetta voi hankaloittaa mittausvirhe ja vastauskato. Merkittävyys taas määräytyy kussakin tilanteessa muista syistä. Voi olla myös niin, että ero joka ei ole tilastollisesti merkitsevä voi olla merkittävä. Sitä ei pidä piilottaa. Jos taas halutaan puhua erosta vaikkapa sukupuolten välillä, on sen oltava myös tilastollisesti merkitsevä. Muuten harhautetaan lukijoita. Valitettavasti media eikä aina tutkijakaan tätä ymmärrä jos tulos tukee hänen omaa näkemystään tai sillä pääsee julkisuuteen. Ole varovainen. Jatkossa tässä osiossa keskityn tilastolliseen merkitsevyyteen.

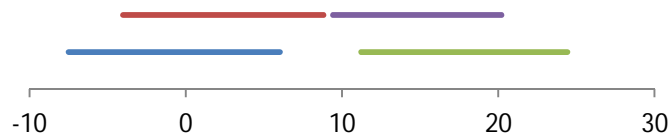
Epävarmuus tilastotieteessä, erityisesti otantaa käytettäessä Aluksi on siis data joka on poimittu otannalla. Sitten on laskettu siitä estimaatti. Koska siihen liittyy **satunnaisuudesta epävarmuutta** ja myös **systemaattista**, on syytä arvioida tämän suuruus. Systemaattisen eli **harhan** osalta se voi olla vaikeahkoa mutta satunnaisen osalta on olemassa mittareita kuten **keskivirhe**, **virhemarginaali**, **p-arvo**, **t-arvo**, ... Tilastolliset ohjelmistot tuottavat näitä ja jos osaat tehdä asiat oikein, saat harhattomia tuloksiakin. Annan kolme esimerkkiä esitystavasta. Muistanet, että 95% luottamusväli = median käyttämä virhemarginaali = estimaatti $\pm 1,96 \cdot$ keskivirhe normaalijakaumaoletuksella.

95%:n luottamusvälit numerollisesti esitettynä neljälle estimaatille A, B, C ja D

A (-4,0; 8,8) B (9,4; 20,2)

C (-7,5; 6,0) D (11,2; 24,4)

95%:n luottamusvälit viivaesityksenä



Luulen että ylemmässä vaihtoehdossa kestää jonkin aikaa nähdä tilastollisesti merkitsevät erot, mutta alemmasta sen näkee heti: jos luottamusvälit eivät mene päällekkäin, ero on merkitsevä. Siis mitkä estimaatit eroavat merkitsevästi?

Jos haluat tulkita tulosta enemmän, niin kerron mitä mikäkin estimaatti tarkoittaa. Kaikki ovat surveyystä jossa oli mahdollisuus vastata joko netissä tai puhelimitse haastattelijan merkitessä vastaukset ylös.

A ja D ovat nettivastauksia, B ja C siis puhelinvastauksia. Toisaalta A ja B vastaajat ovat alle 30-vuotiaita ja C ja D taas 60 vuotta täyttäneitä.

Itse kyselyn idea on selostettu seuraavalla sivulla. Sinulla on siis monia mahdollisuuksia tulkita tulosta. Huomaa että esimerkkiin ei ole otettu kaikkia ikäryhmiä vaan sellaiset jotka havainnollistavat sekä nettivastaamista että puhelinvastaamista ja luottamusväleissä on eroja.

Edellä olevassa esimerkissä (joka jatkuu) on kyse estimaateista ja luottamusväleistä, joiden arvot voivat olla sekä negatiivisia että positiivisia. Toisaalta arvojen jakauma ei ole kaukana normaalista. Kyseessä on kuluttajien luottamusindikaattori josta arvot ovat välillä -100 ja +100 mitä käsittelemme pitkässä kirjaesimerkissä sekä Kansantaloudellisessa Aikakauskirjassa. Indikaattori on muodostettu neljästä muuttujasta joiden arvoväli on juuri tuo ja sen jälkeen on otettu niistä keskiarvo. Tämä ei siis ole meidän harjoitustemme tyypillinen tilanne jossa väli on $[0, 100]$. Tuollainen yhdistemuuttuja on havaittu havainnollisemmaksi luottamusindikaattorille jossa kuluttajien ei tarvitse vastata kuin kolmella kategoriolla:

- Tilanne parantunut ($=+100$)
- Tilanne pysynyt ennallaan ($=0$)
- Tilanne huonontunut ($=-100$).

Mutta asia ei ole aina noin yksinkertainen kuten tässä Helsingin Sanomien kirjoituksessa.

HS 18.2.2011

■ TAUSTA

Mikä on virhemarginaali?

MIELIPIDETIEDUSTELUJEN toinen kummajainen on virhemarginaali.

Perinteisesti tiedustelujen lopussa muistetaan mainita, että tulosten virhemarginaali on suurten puolueiden osalta kaksi prosenttia suuntaansa.

Se on vain peukalosääntö, joka kertoo yli 20 prosentin kannatusta nauttivan puolueen virhemarginaalin kahden tuhannen hengen otoksesta.

Jos esimerkiksi kokoomuksen kannatus olisi gallutuloksen mukaan 20 prosenttia, tosi-

asiassa kokoomuksen kannatus on tällöin kansalaisten keskuudessa 95 prosentin todennäköisyydellä mitä tahansa 18,3 prosentin 21,8 prosentin välillä.

Näin tarkka vaihteluväli kannatukselle on tällöin 3,5 prosenttiyksikköä. Tuon vaihteluvälin sisällä kaikki tulokset ovat yhtä ← todennäköisiä.

Toisin sanoen mainittu 20 prosentin tulos ei edes ole yhtään todennäköisempi kuin mikään muukaan luku 18,3–21,8 prosentin välillä.

Juha-Pekka Raeste

Tämä lause on täysin pielessä.

Tässä kaksi lehtileikettä joissa asia esitetään kylmemmin eli ilman tulkintoja. Tutki selosteiden eroja.

HS 18.3.2011

FAKTA

Näin tutkittiin

- Tutkimus on toteutettu puhelinhaastatteluina 11. – 16. 3.
- Tutkimuksen näyte edustaa maan 18 vuotta täyttäneitä väestöä.
- Haastateltuja on tuhat.
- TNS Gallup Oy on toteuttanut tutkimuksen Helsingin Sanomien toimeksiannosta.
- Tutkimustulosten virhemarginaali on noin kolme prosenttiyksikköä suuntaansa

Huomaa että virhemarginaali on pienempi pienemmille kannatuksille.

Käsite 'suuntaansa' perustuu normaali jakaumaan.

HS 7.7.2012

FAKTA

Näin tutkimus tehtiin

- TNS Gallup selvitti Helsingin Sanomien toimeksiannosta puolueiden kannatusta eduskuntavaaleissa.
- Haastateltavilta kysyttiin myös, kuinka varmoja he ovat puoluevalinnastaan.
- Haastatteluja tehtiin 2 487. Vastaajat edustavat Suomen äänestysikäistä väestöä Ahvenanmaata lukuun ottamatta.
- Tutkimus toteutettiin 4. kesäkuuta – 1. heinäkuuta 2012.
- Tutkimuksen virhemarginaali on suurimmillaan vajaat 2 prosenttiyksikköä suuntaansa.

FAKTA

HS 7.4.2013

Näin tutkittiin

- TNS Gallup toteutti HS:n pyynnöstä tutkimuksen kansalaisten suhtautumisesta vireillä oleviin kansalaisaloitteisiin.
- Tutkimus tehtiin puhelinhaastatteluna 19. 1.– 27. 3. 2013. Haastattelujen määrä oli 1 004.
- Otos edustaa maan 15 vuotta täyttäneitä väestöä Ahvenanmaata lukuun ottamatta.
- Tutkimustulosten virhemarginaali on suurimmillaan kolme prosenttiyksikköä suuntaansa.

Tilanne kehittyi koska lehti ymmärsi tulkinneensa virhemarginaalin väärin: HS 19.3.2011. Syynä oli se, että opetin toimittajaa ja tein paremman graafisen esitystavan. Ks. lehdestä tai Kanava 8/2011. Myöhemmin lehti palasi entiselle tyyliilleen, kun Raeste siirtyi toisiin hommiin lehden sisällä.

TAUSTA

Virhemarginaali on satunnais- ja harhamarginaalin sekasikiö

VIRHEMARGINAALI on epäselvä käsite, joka syntyy satunnais- ja harhamarginaalin yhdistämisestä.

SATUNNAISMARGINAALI mittaa yhdessä tutkimustilanteessa tietyn otosmäärän sisällä olevaa todennäköisyyttä.

Jos esimerkiksi yhdessä kahden tuhannen hengen mielipidetiedustelussa 20 prosenttia vastaajista sanoo äänestävänsä kokoomusta, tutkija tietää taulukosta katsomalla, että kannatuksen satunnaismarginaali on 1,9

prosenttia suuntaansa. Toisin sanoen kokoomuksen kannatus vaihtelee 95 prosentin todennäköisyydellä 18,1–21,9 prosenttiyksikön välillä.

Silti keskimäinen ilmoitettu 20 prosentin luku on todennäköisin, ja noin kaksi kertaa todennäköisempi kuin luku 18,1 tai 21,9. Tässä tapauksessa todennäköisyys noudattaa pitkälti ns. normaalijakaumaa, jonka kansa tuntee Gaussin käyränä.

HARHAMARGINAALI selittää eroa, joka syntyy kun esimerkiksi

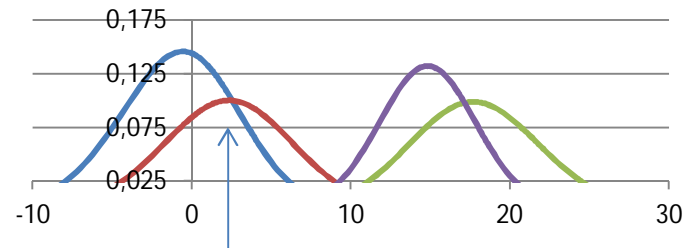
Taloustutkimus ja TNS Gallup tutkivat samalla viikolla puolueiden kannatusta, mutta tulokset ovat aivan erilaisia.

Harhamarginaali on systemaattinen, aineiston keruussa ja laskentametodissa oleva epävarmuus. Harhan mahdollisuus on erityisen suuri uusissa tilanteissa. Esimerkiksi nyt, kun perussuomalaisten gallup-kannatus on lyhyessä ajassa moninkertaistunut, tämä mahdollisuus kasvaa. Sitä ei kuitenkaan voida varmasti arvioida.

Juha-Pekka Raeste

Epävarmuus tilastotieteessä, erityisesti otantaa käytettäessä

95%:n luottamusvälit normaalijakaumallisena käyräesityksenä



Todennäköisyysmassa käyrän alla 95%

Tässä sama kolmannen kerran esitettyä niin kuin se mahdollisimman oikein on. Todennäköisyysmassan koko määrä = 100% ja yleensä normaalijakauman tapauksessa pysytään massan keskellä, ja siten että reunoille jää 2,5% kummallekin puolelle eli keskelle 95%. Tässä on kyse 95%:n luottamusvälistä. Jos haluat jonkun muun, jätä reunoille muu määrä. Kyse on tällöin siitä kuinka suuren riskin rohkenet ottaa. Sitä pienempi, mitä vähemmän jätät reunoille. Älä kuitenkaan aina riskiä pelkää! Oikeaan osumisen todennäköisyys on sitä suurempi mitä korkeammalla on käyrä, ellei ole harhaa.

Suhteellisia frekvenssejä kuten puolueiden kannatuksia estimoitaessa normaalijakauma toimii melko hyvin koska luottamusväli ei helposti mene negatiiviseksi paitsi jos puolueen kannatus on pieni. Näistä puolueista ei olla kiinnostuneita. Joka tapauksessa luottamusväli ei ole oikein laskettuna (logit- tai probitpohjalta) symmetrinen vaan yläraja on estimaatista kauempana kuin alaraja. Monilla tahoilla ei ole välineitä tällaisiin luottamusväleihin. Viimeisellä sivulla on esimerkki Kongon 'sodan' 'liikakuolleista.' Havaitset negatiivisen määrän kuolleita yhdessä estimaatissa (tässä olisi pitänyt käyttää log-muunnosta jolloin negatiivisia arvoja ei voi tulla) . Älä esitä itse sellaisia lukuja missään. 'Lapsikin' ymmärtää ettei se ole mahdollista. Tuollainen tapa lisää vaan pilailua tilastotieteestä. Älä mene sille tielle!

Table 2.1 Excess Deaths in the Democratic Republic of the Congo, 2001-2007: International Rescue Committee and Human Security Report Project Estimates

Period	IRC (best)	HSRP (best)	IRC (low)	HSRP (low)	IRC (high)	HSRP (high)
May 2001-December 2001	418,400	209,200	180,800	29,800	654,500	402,300
January 2002-December 2002	343,200	257,400	120,100	34,300	583,400	497,600
January 2003-April 2004	607,000	101,200	101,200	-404,700	1,112,900	607,000
May 2004-December 2005	735,000	136,600	419,300	-179,700	1,138,100	539,100
January 2006-April 2007	727,000	158,600	522,000	-31,800	1,050,000	371,300
May 2001-April 2007	2,830,600	863,000	1,343,400	-552,100	4,538,900	2,417,300

Data Sources: IRC and Human Security Report Project.

The International Rescue Committee's (IRC) 'best estimate' of excess deaths in the Democratic Republic of the Congo for the period May 2001 to April 2007 is 2.83 million. Using the IRC's survey data, but a more realistic baseline mortality rate, the Human Security Report Project's 'best estimate' of the excess death toll for this period is 0.86 million. In both cases the margin of probable error is large, as indicated by the wide confidence intervals.

?