

Vale, emävale ja tilasto

Tilastojen tutkiminen on kiinnostavaa ja ne voivat antaa paljon arvokasta tietoa esimerkiksi yhteiskunnassa tapahtuvista muutoksista. Usein tutkimuksia ja niihin liittyviä tilastoja esitetään tai tulkitaan väärin – tahattomasti tai jopa tahallisesti. Tässä kymmenen yleistä kompastuskiveä, joiden suhteen kannattaa olla tarkkana.

1. Onko syy- ja seuraussuhde oikea?

**JÄÄTELÖNSYÖNTI
LISÄÄ
HUKKUMISKUOLEMIA!**



	SYÖTY JÄÄTELÖMÄÄRÄ <small>litraa/henkilö</small>	HUKKUMIS- KUOLEMAT
TOUKOKUU	1	0
KESÄKUU	5	10
HEINÄKUU	10	20
ELOKUU	8	15

Kun vedetään johtopäätöksiä kahdesta erilaisesta asiasta kertovista luvuista, on syytä olla tarkkana. Onko asioilla oikeasti joku yhteys, ja onko syy ja seuraussuhde sellainen kuin oletamme? Vaikuttaako taustalla ehkä joku tuntematon tekijä?

Jos yhdistää jäätelönsyömisestä ja hukkumiskuolemista tehdyt tilastot, niin äkkiseltään voisi päätellä, että jäätelönsyönti lisää hukkumiskuolemia. Ei tarvitse olla suurikaan tilastotietäjä ymmärtääkseen, että syy-seuraus-suhde on järjetön tai että molempia selittää todennäköisesti kolmas tekijä eli säätila. Kun ilma lämpenee, jäätelö maittaa ja kansa virtaa uimarannoille ja vetten äärelle, jolloin myös hukkumiskuolemien riski kasvaa.

Voi myös olla, että edes säällä ei ole mitään tekemistä kummankaan kanssa, vaan selittävänä tekijänä onkin se, että heinäkuu on kesäloma-aikaa ja silloin ihmisillä on enemmän aikaa sekä syödä jäätelöä että oleilla uimarannoilla. Tai sitten selitys löytyy jostain ihan muualta.

Tämä keksitty esimerkki on varsin helposti huomattava väärintulkinta, mutta aina näin ei ole. Se, että jotkut asiat kehittyvät samansuuntaisesti tai korreloivat keskenään, ei automaattisesti tarkoita, että niillä olisi syy-seuraus -suhde.

2. Muutoksen suuruutta liioitellaan

**Tappaja-aine löydetty:
PSEUDOMIINI
KASVATTAA
SYÖPÄRISKIÄ JOPA
100%**

Joskus otsikoihin on poimittu shokeeraavia tietoja esimerkiksi jonkun ruoka-aineen terveysvaikutuksista tai -riskeistä. Usein taustalla on ihan vakavasti otettava tutkimus ja riski saattaa todella kohota.

Usein kuitenkin unohdetaan riskitaso, joka voi olla jo alun perin hyvin pieni. Jos jonkin sairauden saamisen riskitaso on vaikka 1/10 000, niin 100 prosentin lisäys riskitasoon merkitsee sitä, että riski on nyt 2/10 000. Toisin sanoen riskin kasvamisen jälkeen kaksi ihmistä kymmenestä tuhannesta sairastuisi syöpään.

Riskitaso on siis edelleen matala, ja on liioittelua puhua tappaja-aineesta.

3. Prosentti ja prosenttiyksikkö eivät ole sama asia

Tämä kohta on sukua edelliselle ja liittyy jonkun muutoksen kuvaamiseen prosenttiluvuilla.

Hippastanin parlamenttivaaleja seurataan silmä kovana: syksyn vaaleissa valtapuolueen asemasta taistelevat Tango-, Jenkka- ja Sambapuolue. Kannatusmittauksia on tehty tammikuussa ja heinäkuussa ja nyt Tangopuolueen pää-äänenkannattaja Tangoviesti uutisoi näyttävästi, että Tangopuolueen kannatus on kasvanut 200 prosenttia. Totta se onkin, mutta todellisuudessa luvut näyttävät tältä:

**TANGOPUOLUEEN
KANNATUS ON
KASVANUT
200%**

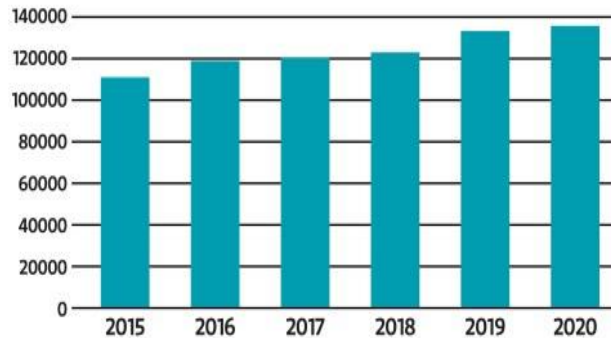
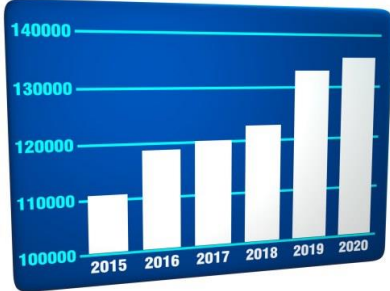
	TAMMIKUU	HEINÄKUU
TANGOPUOLUE	2%	6%
JENKKAPUOLUE	45%	44%
SAMBAPUOLUE	20%	50%

Puolueen kannatus on tosiaan kolminkertaistunut eli kasvanut 200 prosenttia, mutta prosenttiyksikkönä ilmaistuna eli verrattuna puolen vuoden takaiseen prosenttiosuuteen, kannatus on heinäkuussa vain 4 prosenttiyksikköä suurempi kuin tammikuussa. Sambapuolueen kannatus on noussut huomattavasti enemmän: 30 prosenttiyksikköä eli yli 7 kertaa enemmän kuin Tangopuolueen, vaikka Tangoviesti antaakin ymmärtää, että puolueen suosio olisi noussut erityisen paljon.

Prosentti tarkoittaa sadasosaa jostakin. Prosenttiyksikköä käytetään taas prosenttilukuja toisiinsa verrattaessa.

4. Pystyakseli on katkaistu

Hömppäshown katsojaluvut HURJASSA KASVUSSA!



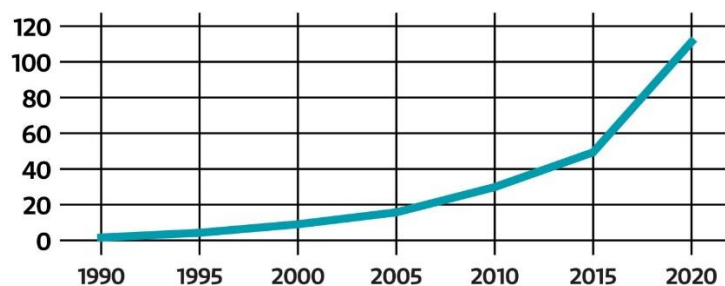
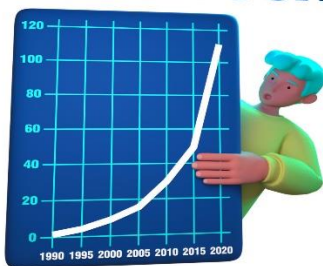
Hömppäshown katsojamäärä ei ole niin hurjassa nousussa, kun sarjan tuotantoyhtiö antaa ymmärtää. Lukujen noususuuntaa liioitellaan yksinkertaisella kikalla: katkaisemalla pystyakseli. Kun kaaviota katsotaan tarkemmin, niin huomataan, että katsojamäärä ei suinkaan ala nolasta, vaan vasta 100 000:sta. Tämän kikan avulla muutos näyttää suuremmalta kuin onkaan.

Kun pystyasteikkoa ei ole katkaistu eli asteikko alkaa nolasta. Huomataan, että katsojalukujen muutos ei olekaan niin suuri kuin edellinen käyrä antaa olettaa. Tällaisissa tapauksissa puhutaan niin sanotusta valekertoimesta. Tässä se on sitä suurempi, mitä ylempää pystyakseli on katkaistu.

Visuaalisella esitystavalla on siis suuri vaikutus siihen, miten pieneltä tai suurelta joku muutos tai kehityskulku näyttää. Myös kahdessa seuraavassa kohdassa on kysymys samantapaisesta manipuloinnista.

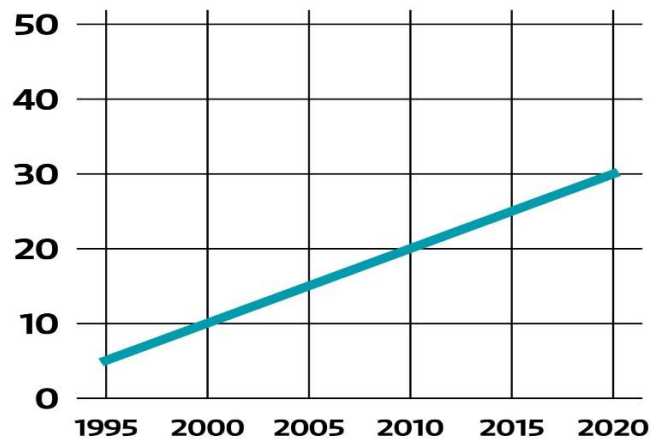
5. Kaavion muotosuhdetta on muutettu

Itä-Touhulan työttömyys RAKETTIMAISESSA NOUSUSSA!



Itä-Touhulan työttömien määrä on toki nousussa, mutta ei niin rajusti kuin kaaviosta voisi ymmärtää. Tällä kertaa tekniikkana on, että kaavion muotosuhdetta on muutettu: sitä on venytetty pystysuunnassa. Näin muutosta saadaan liioiteltua – tai vähäteltyä, kuten allaolevassa kuviossa, jossa muotosuhdetta on venytetty vaakasuunnassa. Molemmissa kuvioissa ovat siis aivan samat tiedot, vaikka käyrät näyttävätkin erilaisilta. Taulukkolaskentaohjelmilla kaavion muotosuhde muuttuu helposti – vaikka tahtomatta – jos kaaviota venyttää pysty- tai vaakasuunnassa.

6. Aika-akseli ei ole tasavälinen



Tämä keksitty kaavio kertoo tuottavuuden kehityksestä, mutta se antaa sen kasvusta vähättelevän kuvan. Kikkana on käyttää kuviota, jossa aika-akseli ei ole tasavälinen. Kaavion alkupäässä pystyviivojen väli on 10 vuotta, lopussa se on vain yksi vuosi. Näin käyrä saadaan näyttämään sellaiselta kuin halutaan. Tasavälinen kaavio näyttää tältä. Pystyviivojen väli on nyt aina 5 vuotta. Kaaviossa on samat luvut kuin edellisessä kaaviossa. Nyt huomataan, että tuottavuus onkin oikeasti kasvanut lineaarisesti.

7. Keskiarvo ei aina kerro kaikkea eikä edes tarpeeksi

**Humppalandiassa
eletään leveästi:
KESKITULO
109 000
EUROA
VUODESSA**

Keskiarvoa käytetään usein kertomaan jonkun asian tai ominaisuuden määrästä jossakin joukossa. Voidaan esimerkiksi sanoa, että aikuinen norsu painaa keskimäärin 3000 kiloa, jolloin luvusta on helppo päätellä, että ihan pieniä eläimiä ei aikuisessa norsulaumassa löntystä.

Aina keskiarvo ei kuitenkaan ole yhtä valaiseva, sillä saattaa olla, että mitattava asia tai ominaisuus on polarisoitunut ja keskittynyt akselin ääripäihin. Silloin keskiarvo antaa väärän kuvan todellisesta tilanteesta.

Otetaan esimerkiksi pieni Humppalandian maa, jossa on vain sata asukasta. Jos Wikipedia kertoo, että Humppalandian keskitulo on 109 000 euroa vuodessa, niin

helposti ajattelemme että maa on vauras ja hyvinvoinnin taso korkea.

Näin toki onkin, jos tulot jakautuvat tasaisesti ja kaikki ansaitsevat suunnilleen saman verran. Saattaa kuitenkin olla, että kääpiövaltiossa onkin 10 hengen kuningashuone, jonka jäsenistä jokainen ansaitsee miljoonan vuodessa: muut asukkaat joutuvat kitkuttelemaan 10 000 euron vuosituloilla. Humppalandia onkin tällöin maa, jossa varallisuuserot ovat erityisen suuria ja suurin osa kansasta elää varsin vaatimattomasti.

Keskiarvoa paremmin Humppalandian tulotasosta kertoisi esimerkiksi mediaani. Mediaani saadaan, kun havaintoarvot asetetaan suuruusjärjestykseen ja otetaan niistä keskimäinen tai kahden keskimäisen arvon keskiarvo. Humppalandiassa tulojen mediaani olisi 10 000 euroa.

8. Nettikyselyä väitetään tutkimukseksi

**Tutkimuksemme
yllättävä viesti:
60%
SUOMALAISISTA
poistaisi autoveron!**

Mediat ja erilaiset järjestöt tekevät usein kyselyjä nettisivuillaan. Joskus niiden kerrotaan olevan tutkimuksia tai kyselytutkimuksia. Kyseessä ei kuitenkaan ole oikea tutkimus, vaan kysely, jonka perusjoukkoa ei tunneta. Sen tuloksia ei voi yleistää koskemaan kaikkia suomalaisia, koska esimerkiksi kyselyn tehneen median lukijat tai järjestön jäsenet ovat valmiiksi valikoitunut joukko.

Oikea kyselytutkimus perustuu edustavaan otokseen kaikista suomalaisista. Yleensä niissä vastaajien määrä on tarkoituksesta riippuen vähintään tuhat.

Vaikka nettikyselyyn osallistuisi useampikin kuin tuhat vastaajaa, ei se silti tee siitä sen parempaa. Vastaajat eivät edelleenkään ole edustava otos suomalaisista, vaan vinoutunut näyte.

9. Ei ymmärretä käsitteitä ja käytetään niitä väärin

**YLI NELJÄNNES
NUORISTA
MIEHISTÄ
VAILLA TYÖTÄ**

Yksi tiedotusvälineissä toistuvasti esiintyvistä virheistä on nuorisotyöttömyyden yliarviointi. Tässä ongelmana on, että ei ymmärretä, mitä käsite nuorisotyöttömyys tarkoittaa. Se ei ole osuus kaikista ikäluokan nuorista, vaan vain työmarkkinoiden käytettävissä olevista nuorista.

Jos siis Tilastokeskuksen työvoimatutkimuksen mukaan 15–24-vuotiaiden miesten työttömyysaste oli 25,6 %, ei se tarkoita, että yli neljännes nuorista olisi työttömänä. Tämä johtuu siitä, että puolet 15–24-vuotiaista miehistä on työvoiman ulkopuolella, esimerkiksi opiskelemassa tai armeijassa.

Esimerkiksi maaliskuussa 2018 ikäluokkaan kuului 319 000 miestä, joista 39 000 luokiteltiin työttömiksi. Työttömiä nuoria miehiä oli siis todellisuudessa vain 12 prosenttia ikäluokasta.

10. Kausivaihtelua ei oteta huomioon

**KUUKAUDESSA
30 000
UUTTA
TYÖTÖNTÄ**

Tieto pitää periaatteessa paikkansa, kun verrataan huhtikuuta ja toukokuuta 2018. Tällä kertaa ongelmana on, että vertailussa ei oteta huomioon kausivaihtelua. Yleensä työttömien määrä vuosittain on nimenomaan toukokuussa kaikkein suurin, koska silloin työmarkkinoille tulee uusiksi työnhakijoiksi paljon opiskelunsa päättäneitä ihmisiä.

Oikea tapa olisi verrata työttömien määrää edellisen vuoden toukokuuhun, jolloin huomataan, että itse asiassa vuodessa työttömien määrä on laskenut 35 000 henkeä.

Toinen mahdollinen oikea tapa tutkia työttömyyden trendiä on verrata vuosineljänneksiä keskenään. Tällöin kausivaihtelun vaikutus ei ole niin suuri. Tässäkin tapauksessa huomataan, että työttömien määrä oli vuoden 2018 toisella neljänneksellä itse asiassa laskussa – ei nousussa.

Kausivaihtelun huomioonottaminen on työttömyydestä puhuttaessa tärkeää, sillä työttömyydessä on paljon luonnollista kausittaista vaihtelua eri aloilla. Esimerkiksi rakennusala työllistää enemmän kesällä, Lapin turismi taas enemmän talvella. Kausivaihtelu näkyy myös muun muassa teollisuustuotantoa, vientiä ja tuontia käsittelevissä tilastoissa.

Jutun pääasiallisina lähteinä on käytetty kalvosarjoja Tulosten visuaalinen esittäminen eli tilastografiikan perusteita (Vesa Kuusela) ja kirjaa Tilastot käyttöön! Opas tilastojen maailmaan (Jussi Simpura & Jussi Melkas, Gaudeamus 2013).