Suuremad vead ja eksimused statistiliste andmete tootmisel ja ajakirjanduses kasutamisel

Peeter Annus

Statistikaameti peaanalüütik

peeter.annus@stat.ee

eesti statistika

MILLEST TULEB JUTTU

- Statistilised andmed on teaduse jaoks pigem toormaterjal. Algul jälgin **etapiti andmete tootmise käiku,** kuidas kujuneb statistilistele andmete kontekst. Igas etapis **toon esile seal tehtavaid vigu.** Igast etapist saavad andmed kaasa mõjutusi, mis pärandina kas nähtavalt või peidetult andmetega levisse ja tarbimisse kaasa lähevad.

- Järgnevad mõned näited Statistikameti statistikakommunikatsiooni kogemusest ja ametliku statistika loodud lisaväärtuse valem.

- Kui toormaterjal levitamise etappi jõuab, siis tekib statistilise materjali tarbijal vajadus hinnata andmete kasutamiskõlblikkust. Võimalike hindamisnäitajate kirevus on suur, **pakun** ettekande viimases osas **kolme lähenemist** kuidas neid näitajaid, kvaliteedi nõudeid, tüüpilisi eksimusi võiks **ülevaate saamiseks korrastada**.

Kommunikatsiooni probleemide tagamaa

- **Milline peaks olema ametliku statistika lõplik väljund?**
- Ajalooliselt olid statistikat tootvad asutused mõeldud teenindama väiksearvulist ja mõjukat tarbijat (valitsusasutused, akadeemilised eksperdid jne.) ning samu tooteid võis siis kasutada ka ülejäänud elanikkond.
- Tagajärg: tootja ise andmeid ei analüüsinud, vaid andis need üle tarbijale. Sellise praktika tagajärg oli, et metoodika tutvustuse laad ja keelekasutus **sobisid küll ekspertidele, kuid mitte laiemale üldsusele.**
- Algselt oli andmete põhiline kvaliteedikriteerium täpsus ja õigeaegsus, mis olid ka usaldusväärsuse aluseks
- Aegapidi on hakanud kasutajate ringi laienema. Kvaliteedinõuetest lisandusid: kättesaadavus, arusaadavus ja võrreldavus – toetuseks tarbijale, kes tehnilisi peensusi ei valda. Nüüd on aktuaalne: huvitav esitus ja veenvus – üldsuse seas levitamiseks tuleb alluda tururee glitele.
- SA on valiku ees statistiliste materjalide levitamisel: kas tuua esile ainult olulist või otsida ka huvitavaid fakte. Probleem: esiletoodud oluline ei ole tingimata huvitav.

Kuidas statistilisi andmeid toodetakse

- Võtan statistika tootmistsükli läbi sammsammult. Igas sammus vaatlen vigade tekkimise võimalust. Statistika tootmine ongi vigadega võitlemine, seetõttu on hea teada millest vead tekivad.
- Andmete tootmise käiku jälgides selgub statistiliste andmete konteksti kujunemine. Kujunemist aitab täiendavalt mõista võrdlus teiste kanalitega, kus arvandmeid on tekitatud või kasutatud.

- Valisin Statistikaameti - muuseas, erinevalt paljudest riigiasutustest on SA sisuliselt tootmisettevõte – kõrvale võrdluseks järgmiste infoäri esindajate tegevuse …

- **Turu-uuringu firmad**, nende tegevus on sarnane SA omale.
- Ajalehtede veebiportaalides **Gallupiks** nimetatult üles pandud küsimuste ja nende vastusevariante abil andmekorje. Selle tulemust võib nimetada ka statistikalaadseks tooteks.

ETAPP 1 Lähteülesanne ja küsimustiku koostamine

SA: on riiki esindavate institutsioonide tellimuste täitja ja kogu ühiskonda kirjeldava statistika põhiline tootja. Tarbijate vajadusi kogutakse igaaastaselt ja tööde plaani kinnitab Vabariigi Valitsus.

Võimalikud vead ja allikad

- M**õni oluline näitaja jääb uuringust välja**.
- **Mõni oluline teema jääb katmata.** Ametkondade vahelised teemad; uued üldriiklikud teemad. Teemad millel ei ole konkreetset tellijat. Enamasti ei jätku ressursse. Protseduurid pole piisavalt paindlikud.
- Milliste näitajate kogumine on vajalik riigi jätkusuutlikuks juhtimiseks. Mida mõõta, et protsessid oleks juhitavad? Riigi seisukohalt on tüüpilised SKP, keskmine palk, töötuse määr, tarbijahinna indeks ..
- **Mõne küsimuse sõnastus ei ole asjakohane** Puudub ressurss, et kõiki küsimustikke ja küsimusi testida.
- **Võrreldavus üheks otstarbeks võib takistada andmete võrdlemist muu olulise andmebaasiga.** Tüüpiline mõjutaja on Eurostat, mille tingimustel tehakse ca 85% SA statistikatöödest.

Võrdluseks infoäri ..

- **G:** üks küsimus aktuaalsel teemal lugeja huvi äratamiseks. Küsimuse kvaliteet on kõrvalise tähendusega. Rõhuasetus on meelelahutusel, artiklile lisatud küsimus aitab jutustada selle lugu, kaasates lugeja.
- **TU** paindlikum kui SA: küsimustik sünnib koostöös tellijaga, uuringu eelarve raames. Võrreldavus kokkuleppel.

ETAPP 2 Valiku-uuringu valimi koostamine

SA: koostab küsitletavate valimi erinevate registrite andmete põhjal. Valimid on suured, 10 000 järgus. Suurem valim lisab täpsust ja võimaldab ka väikeseid valdkondi või sihtrühmi võrrelda.

Võimalikud vead ja allikad

- **Valim on liiga väike** usaldusväärseks üldistuseks üldkogumile. Standardviga ja usaldusvahemik jääb liiga laiaks. Pilootuuringuks on väike valim piisav, kuid kui need tulemused satuvad avalikkuse ette, siis tekib valetõlgenduse oht . Piiravaks on ressursinappus.

- **Valim ei ole esinduslik**. Kõik olulised sihtrühmad ei ole valimis või on proportsioonist väljas. Üldkogumi kontaktinfo ei ole täies ulatuses kättesaadav või pole teada sihtrühmade täpsed proportsioonid. Samas väikerühmade liikmeid võetakse valimisse nende osakaalust rohkem piisava usaldusvahemiku tagamiseks.

Võrdluseks infoäri ..

- **G**: valimit ei määratleta, kõik portaali külastajad on potentsiaalsed vastajad .
- **TU**: Suurus ja koosseis on kokkuleppel kliendiga. tüüpiliselt 300–1000, äriturul väiksem, eraisikute turul suurem valim.

ETAPP 3 Küsitlus

SA: Riikliku statistika seadus sätestab SA-i õiguse andmeid koguda, ettevõtetel on vastamine kohustuslik. SA küsitlejaid peetakse veidi usaldusväärsemaks kui ärifirmade küsitlejaid.

Võimalikud vead ja allikad

- **Mittevastamisest tingitud viga** tekib küsitlemisel alati, sest 100% valimist ei vasta ja mittevastajate asendamisel ei ole vastavus täpne. Viga ei ole mõõdetav, kuid on (osaliselt) kompenseeritav kaalumise ja imputeerimise abil.
- **Mõõtmisviga** tekib andmete kogumisel, kui vastaja on andnud ebatäpseid andmeid. Mõõtmisviga saab kaudselt hinnata. Vea tekitajaid: vastaja mõistab küsimust valesti, vastaja ei suuda vastust leida või moonutab teadlikult vastust; küsitleja eksib küsitlusmetoodika vastu.

Võrdluseks infoäri ..

- G: portaali külastajad valivad ise, kas vastavad. Portaali külastamine on juba eelnev valik. Rakendatakse mitmekordselt vastamise vältimise meetmed.
- TU: samad vead kui SA-l. Valimid on väga erineva koosseisuga, seetõttu võib nende vigade mõju hindamine olla keerukam.

Uued andmeallikad. Trendiks on küsitluse asendamine registriandmete hõlvamisega, tulemuseks : valimiviga taandub, kuid töötlusviga suureneb, definitsioone ei saa ise valida ja seetõttu on vaja andmete kohandamiseks töötlusalgoritme.

ETAPP 4 Andmetöötlus

SA: sellel etapil püütakse parandada eelmiste etappide vigu, tegevus on väga töömahukas.

Võimalikud vead ja allikad

- **Töötlusviga** võib tekkida andmete kaalumise, kodeerimise, imputeerimise jne käigus metoodiliselt ebasobivate võtete rakendamisel, aga lihtsalt ka sellest, et tühikute täitmisel on täpselt teadmata, kui kaugele jääb asendväärtus tegelikust väärtusest.

Võrdluseks infoäri ..

- G: vigu ei kompenseerita.
- TU: Andmetöötlus on pigem valiv ja vigu püütakse kompenseerida väiksemas mahus kui SA-s. Tööde maht sõltub kokkuleppest kliendiga.

ETAPP 5 Andmete kirjeldamine/analüüs

SA: põhiliselt esitatakse andmed mitmemõõtmeliste tabelite kujul, samuti arvutatakse indekseid ja keeruka arvutusvalemiga koondnäitajaid nagu SKP. Nende põhjal koostatakse pressiteateid, blogipostitusi, väljandeid.

Võimalikud vead ja allikad

- **Valikuviga** on määratud valimi suurusega ja see määrab usaldusvahemike laiuse valimist saadavate hinnangute laiendamisel üldkogumile. Vea suurus on täpselt arvutatav.
- **Visuaalsed moonutused** tulemuste graafilisel esitusel. Eriti sõltuvalt skaalade valikust.
- **Ekslikud statistilised järeldused -** olulise erinevuse, põhjusliku seose jne. olemasolu või puudumine.

Võrdluseks infoäri ..

- G: tulemus on tavaliselt esitatud graafiliselt, vastaja saab hinnata enda asendit teiste vastanute suhtes.
- TU: Väiksemate valimite tõttu on valikuviga tavaliselt suurem kui SA-il. Tänu tihedamale tagasisidestusele kliendiga on suurem tõenäosus võimalike eksituste avastamiseks.

SA toodete ja teenuste levi, 2015

- Meediakanalid kajastasid Statistikaametit ja riiklikku statistikat 7262 korda - neist 3503 käisid pressiteadete ja 441 blogipostituste kohta. Meediakajastuste koguarvust ligi poole andsid pressiteadete põhjal ilmunud kajastused.

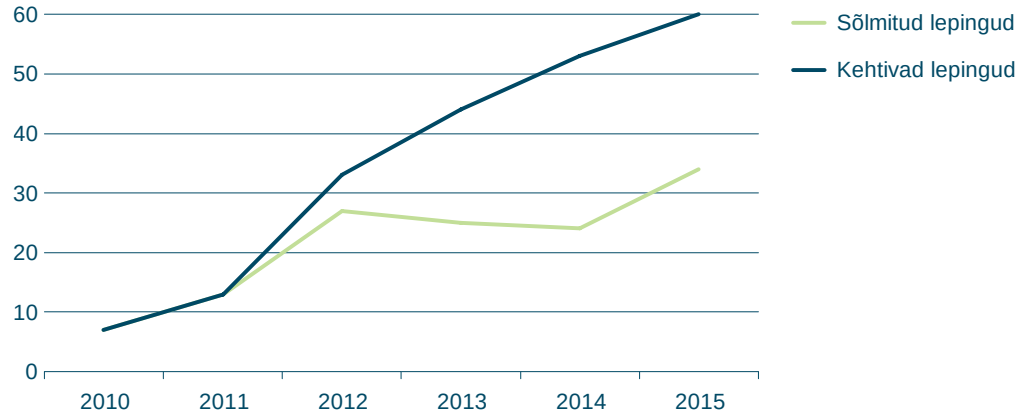
- Statistikaamet avaldas **138 pressiteadet,** neid **kõiki kajastati meedias**, igat pressiteadet **keskmiselt 25 korda**. Enim huvitus meedia majanduskasvust ja palgaturgu pressiteadetest.

- Veebilehe külastuste arv nädalas: *ca* 11 000 külastust, millest 82% Eestist.
- Blogikülastusi aastas 132 000 korda
- Väljaandeid laeti veebist alla 27 000 korral.
- Esitati 2880 teabenõuet.

Statistika üksikandmetele juurdepääsu võimaldamine teadustööks

- Teadusasutustel on võimalik saada teadustööks juurdepääs Statistikaameti statistikatöödes kogutud üksikandmetele.
- Juurdepääsu võimaldamine on igal konkreetsel juhul Statistikaameti kaalutlusotsus, mille tegemisel arvestatakse taotletavate andmete alusel isiku tuvastamise riski ja andmete konfidentsiaalsust.

Üksikandmete kasutamise lepingud, 2010–2015



STATISTIKA KOMMUNIKATSIOON
Näide1: sama teema, erinevad rõhuasetused pealkirjas

- Tootjahinnaindeks septembris kaks protsenti mullusest madalam **Postimees**

- Tööstuse tootjahinnaindeks tegi väikse tõusu **Äripäev**

Statistikaameti pressiteatest ….

- Tööstustoodangu tootjahinnaindeksi muutus oli septembris võrreldes augustiga 0,2% ja võrreldes eelmise aasta septembriga -2,0%, teatab statistikaamet.

10.2014

STATISTIKA KOMMUNIKATSIOON
Näide2: sama teema, erinevad rõhuasetused pealkirjas

- **„Vabade töökohtade arv kasvas märkimisväärselt“** **Postimees**
- **„Tööjõupuudus süvenes järsult“** **Äripäev**

Statistikaameti pressiteatest ….

- Eesti ettevõtetes, asutustes ja organisatsioonides oli 2016. aasta I kvartalis ligi 8300 vaba ametikohta, teatas statistikaamet. Vabade ametikohtade arv suurenes eelmise kvartaliga võrreldes 25% ja 2015. aasta I kvartaliga võrreldes 14%. ……
  juuni 2016

STATISTIKA KOMMUNIKATSIOON
Statistiliste materjalide levitamine

- Statistiline info avaldatakse, mis saab edasi?

- Enesekehtestamine meediaringluses eeldab, et statistilised materjalid on veenvad ja köitvalt esitatud. Levi edukust mõjutab ka statistikatootja kuvand meedias. Samas väheneb statistikatootjale endale oluliste kriteeriumite väärtus.

- Teisalt, statistilistel andmetel endil pole väärtust kui neid keegi ei kasuta. Kui andmeid kasutatakse, siis vajavad nad tõlgendamist ja kui statistika tootja ei ole neid tõlgendanud, siis teevad seda andmete kasutajad ise. Tõlgendus peaks siis olema veenvalt ja huvitavalt esitatud. Tavaliselt ei ole eriti huvitavalt.

- Mida siis teha, et statistilised materjalid saaks laiema leviku?

STATISTIKA KOMMUNIKATSIOON

VAS = ametliku statistika loodud lisaväärtus

VAS = { N * [(QSA * MF) * RS * TS * NL]} – CS

- N = huvitatute arv
- QSA = toodetud statistilised andmed
- MF = meedia roll
- RS = statistiliste andmete asjakohasus
- TS = ametliku statistika usaldusväärsus
- NL = kasutajate statistiline kirjaoskus
- CS = statistiliste andmete tootmiskulud
  - *Valem ei ole konkreetse arvutuse tegemiseks. Siiski on oluline märgata, et kui korrutise mõni tegur on 0, siis lisaväärtust ei teki, jäävad vaid kulud*

Allikas: Enrico Giovannini endine OECD peastatistik

Kes räägib tõtt?

Peeter Raidla

JAGA FACEBOOKIS

KOMMENTEERI PRINDI ARTIKKEL JAGA ARTIKLIT 25. august 2016 09:06

Kui otsida andmeid töötute arvu kohta Eestis, siis ei ole pealkirjas esitatud küsimus sugugi pelgas sõnakõlks.

Statistikaameti kommentaar Maa Elu juhtkirjale „Kes räägib tõtt?" (vt 25.08.2016 Maa Elu)

Kaja Sõstra
Statistikaameti metoodika ja analüüsi osakonna juhataja

JAGA FACEBOOKIS

Artiklis on võrreldud Statistikaameti tehtud riiklikku statistikat ja registrite statistikat. Erinevused ei tähenda, et üks on õige ja teine vale – metoodikad on erinevad.

STATISTIKA KOMMUNIKATSIOON
Eelneva näite sisu selgitus: töötute arv

- Töötukassa andmeil oli 27 209 töötut, Statistikaameti andmeil 45 300 tööotsijat. *06.2016*

- Näitajatel on sarnane nimetus, kuid erinev tulemus. SA tööjõu-uuring saab arvu elanikkonda küsitledes – need kes ei tööta, aga otsivad tööd. Töötukassa avaldab registreeritud töötute arvu. Kõik töötud ei registreeri ennast Töötukassas, vaid otsivad muid võimalusi tööotsimiseks.

KUIDAS ORIENTEERUDA STATISTILISTES ANDMETES

Statistiliste materjalide levitamisel tekib vajadus **hinnata andmete kasutamiskõlblikkust**. Sõltuvalt otstarbest saab hinnata erineva põhjalikkusega. Pakun siin kolme lähenemist …

A. Küsida **iga kvaliteedinõude kohta kontrollküsimusi**. Kuid neid küsimusi on ülekoormavalt palju.

B. Küsida üldisemalt kui **tugevad on tõendid** andmete kvaliteedi kohta.

C. Kontrollida, kas andmetes pole tüüpilisemaid **moonutusi** või kus peaks lihtsalt **ettevaatlik olema**

A. Kvaliteedinõuete kontrollküsimused 1

Euroopa statistikasüsteemis on viis kvaliteedi põhimõõdet:

… asjakohasus; täpsus ja usaldusväärsus; ajakohasus ja õigeaegsus; võrreldavus; sidusus. Igaüks neist jaguneb omakorda üksikasjalikumateks mõõdikuteks. Näiteks mõõdikutest on lisatud mõned kontrolliküsimused iga mõõtme juurde.

- ASJAKOHASUS Kas andmed on vajalikud, kas info on piisavalt üksikasjalik ja sobivate sihtrühmade kohta?
- TÄPSUS mida on teada erinevate vigade kohta - valikuvea suurus, milline % esialgsest valimist jättis vastamata; kuivõrd järgiti küsitluse metoodikat ? Kui selged olid kasutatud mõisted ja küsimuste sõnastus? Mida ütleb kvaliteediraport?

A. Kvaliteedinõuete kontrollküsimused 2

- ÕIGEAEGSUS Kas andmed ilmusid kavandatud ajal? Kas võib järgneda täpsustusi? Kui värsked on andmed? Kas on tegemist kordusuuringuga?

- VÕRRELDAVUS ja SIDUSUS Kas mõistete määratlused ja küsimuste sõnastused on võrreldavates andmebaasides samad? Kas on võimalik moodustada pikaajalisi andmeridu? Kas mingi välissündmus mõjul ei ole võrdlemine enam võimalik?

➢ *Mitmed kvaliteedinõuded on piirangud tootmisprotsessile tarbija huvides ja tootja teeb kompromisse tootmiseks vajalike ja tarbijale vajalike nõuete vahel.*

B. Tõendite tugevuse kontrollküsimused

- **Andmete allika kontrollimine**. Kes, millal ja millise probleemi lahendamiseks andmed tekitas?

- **Kas andmed on esitatud kontekstis, millises kontekstis**? Valimi suurus ja esinduslikkus, andmete korje meetod. Mõistete selgitus.
- Kas olukorra mõistmiseks neist **andmetest piisab**? Siiski, kõike ei ole võimalik küsida.
- Kas **järeldused** tulenevad esitatud statistilistest andmetest?

- Kuivõrd **ootuspärane** on huvipakkuv tulemus.
- Mida näitavad **võrdlusandmed** või teised tõendid.

- Millist **vaatenurka** püüab statistilise info esitaja rõhutada? Ka erapooletul tõlgendamisel on võimalik rohkem kui üks valik.

- Kas on püütud **tahtlikult** tulemusi **kallutada?**

C. Kus olla ettevaatlik 1

Metoodiliste mõistete ja uuringu sisu mõistete tähendused näiteks ….

- Vastaja mõistab ekslikult küsimuse sõnastust, seetõttu on saadud tulemuse tõlgendus ekslik
- Kasutatakse mõistet „keskmine“, täpsustamata, millist tüüpi keskmisega on tegemist
- Sama sõna tähistab erinevates uuringutes erineva sisuga mõisteid.
- Sama sõna tähendus uuringus ja tavakasutuses võib olla erinev. Nt. vastavalt sõna tavatähendusele on tõlgendatud riikide pingerida õnneuuringutes.

Järelduste korrektsus näiteks ….

Andmete esitusviis viib valejäreldustele

- Visuaalsed moonutused arvjoonistel
- Väga suurte/väikeste arvude korral esitatakse ainult protsente, kuid vaja oleks absoluutarve taustaks
- Statistilisi andmeid on tekstis kasutatud kuid seal esitatav väide või seisukoht neile ei toetu

Andmetest tehakse ekslikke statistilisi järeldusi

- Statistiliselt oluline erinevus võrdluses ei pruugi tingimata tähendada eluliselt olulist erinevust.
- Korrelatsioonid ei näita põhjuslikkust

C. Kus olla ettevaatlik 2

Uuritud valimilt üldkogumile üldistamine , näiteks ….

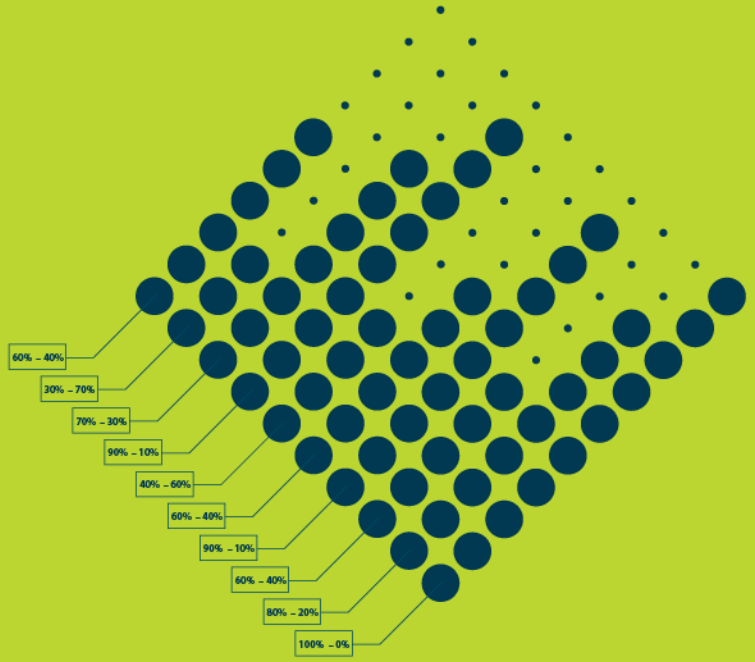
- Sihtrühma ebapiisava suuruse korral
- Juhul kui valim ei ole esinduslik

Andmete võrdlemine , näiteks …

- Kui puudub võimalus selgitada kas võrreldavad andmed vastavad ühtsetele reeglitele
- Metoodika või olud on kordusuuringul muutunud

Andmete piisavus, näiteks …

- Kui andmed on poolikud, olulised andmed puuduvad, selle põhjusi ei leia.
- Kui andmete tõlgendus on selgelt kallutatud rõhuasetusega



eesti statistika