

Vead ja täpsus statistikas

On olemas kolme tüüpi valesid: valed, alatud valed ja statistika.
-Disraeli

Tõepoolest, kasutades statistilisi meetodeid aru saamata nende sisust või siis, halvemal juhul, arvestades kuulajate/lugejate asjatundmatust, on statistika abil valet vanduda küllalt lihtne. Paraku tuleb tõdeda, et seda “võimalust” tõepoolest lubamatult tihti endalegi märkamatu(?) kasutatakse. Kuid kas selles on õige süüdistada statistikat või kvantitatiivseid uurimismeetodeid laiemalt?... Vaevalt küll ... igasugune vale jääb ikka esitaja südametunnistusele ehk parafraseerides üht tuntud käibetõde – meetodi mittetundmine ei vabasta vastutusest.

Millised on siis põhilised vead, mida oma tulemuste esitamisel ja interpreteerimisel tehakse? Selle üle tasuks natukene mõtiskleda kahel põhjusel: esiteks, et hoiduda ise valet vandumast ning teiselt, et osata kriitiliselt hinnata teiste poolt esitatud väiteid ehk, et mitte lasta ennast ninapidi vedada (tehtagu seda siis tahtlikult või enese teadmata).

Mugavuse ja ülevaatlikkuse mõttes olen jaganud tüüpilised eksimused kuude suurde rühma:

- Arvutusvead
- Ebasobiv või tulemusi moonutav esitlusviis
- Ebasobiv analüüsimeetod
- Osaline tulemuste kirjeldus e ebapiisav info tulemuste õigeks interpreteerimiseks
- Analüüsitulemuste väär interpreteerimine
- Tulemuste üldistamisel tehtavad vead

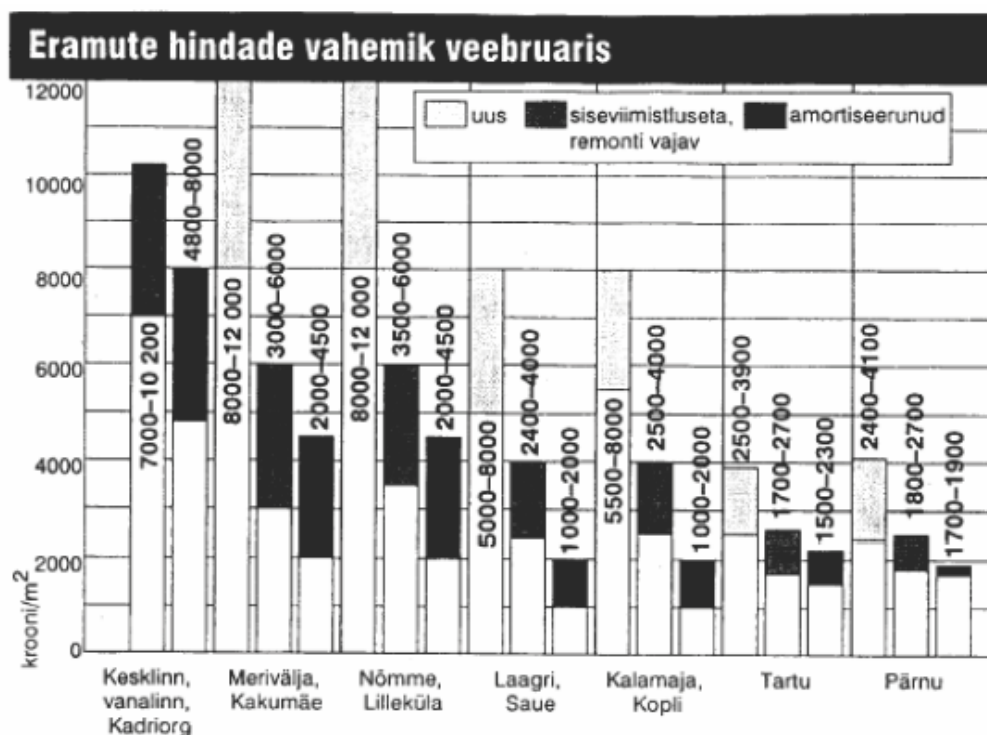
Arvutusvead

Võiks arvata, et arvutusvead, mis mõni aeg tagasi suureks mureks olid, kuna tõsisemate statistikute leidmiseks oli vaja käsitsi või kalkulaatori abil toimetada keerulisi arvutusi tohutu suure hulga algandmetega, on seoses personaalarvutite kasutuselevõtuga kadunud. Mõneti on see tõsi – arvuti ei eksi korrutamisel või jagamisel isegi kui tegu on väga paljude suurte ja keeruliste arvudega. Samas tuleb alati meele pidada, et arvuti teeb täpselt seda, mida kasutaja tal teha käsib ja kui kasutaja käsu andmisel eksib (seetõttu, et ta ei tunne piisavalt hästi vastavat programmi ja/või kasutatavat statistilist meetodit), siis on tulemus suure tõenäosusega vigane. Järele mõeldes võib tõdeda, et esmapilgul absurdsena tunduv väitel nagu oleks kasutajasõbralike statistikapaketide kasutuselevõtt oluliselt suurendanud vigaste tulemuste esitamist, võib oma valus tõetera sees olla.

Ebasobiv või tulemusi moonutav esitlusviis

Kes natukenegi rohkem statistiliste andmete esitlemisega kokku on puutunud teab ilmselt hästi vana käibetõde, et “tabel on arvude surnuaed” ja selleks, et oma tulemusi

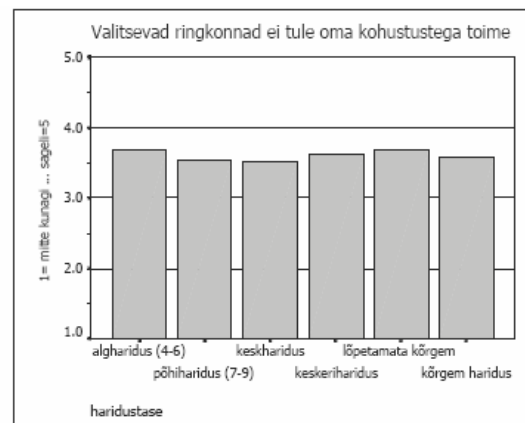
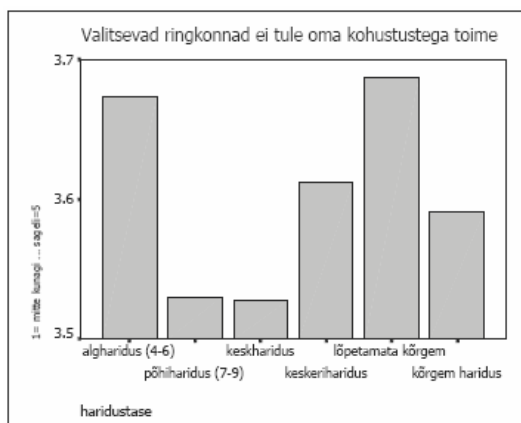
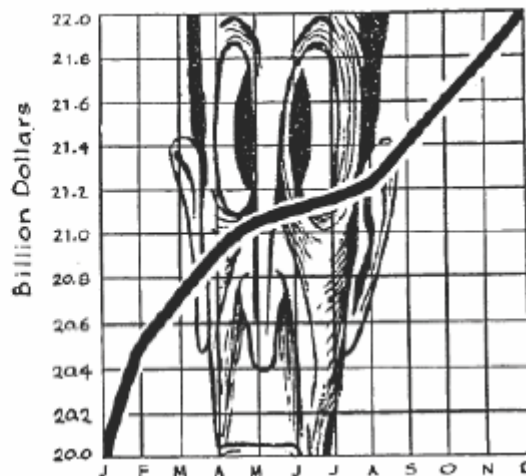
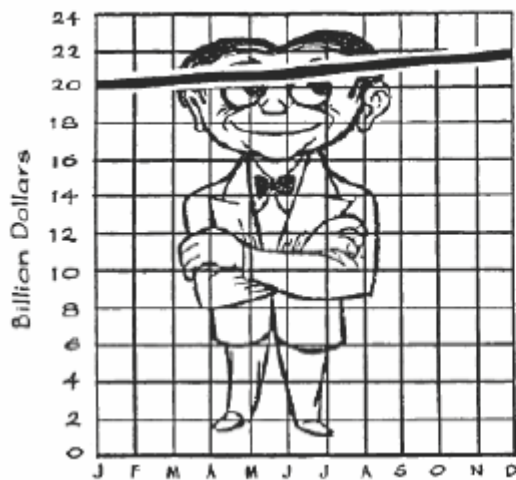
mõjuvalt esitada on vaja graafikut e diagrammi. Selle juhtnööri tingimusteta järgimine võib aga viia üsna kentsakate tulemusteni, eriti kui sinna lisandub veendumus, et veelgi mõjuvamaks annab esitlust muuta siis kui kõik arvuti poolt pakutavad efektid (ebatavaline diagrammi tüüp, kirjud taustapildid, kolmemõõtmelised jne jne) ära kasutada. Parem juhul muutub diagramm lihtsalt nii kirjuks, et sisu enam kuidagi välja lugeda pole võimalik (vt näiteks joonis 1). Halvemal juhul, kui puudub arusaamine erinevate diagrammitüüpide otstarbest ning sisemisest loogikast võib tulemus olla lausa arusaamatu ja eksitav (vt näiteks joonis 2). Kõige halvemal juhul on aga vaatajale võimalik diagrammi abil luua täiesti vale ettekujutus kirjeldatavast nähtusest või protsessist moonutades joonisel olevaid proportsioone (vt näiteks joonis 3). Kui esimesel kahel juhul vaataja tavaliselt tajub, et diagrammiga ei ole kõik päris korras ja oskab ehk seetõttu ettevaatlik olla, siis viimasel juhul on formaalselt kõik just nagu korras ja seega vaatajal ennast eksitava informatsiooni eest kõige raskem kaitsta.



Joonis 1. Näide diagrammist, mis on lihtsalt liiga kirju selleks et seal sisalduvat infot oleks võimalik välja lugeda.



Joonis 2. Näide diagrammist, kus arvandmete ebasobiva paigutusega tulpadele on olemuselt lihtne info muudetud arusaamatuks.



Joonis 3. Näide diagrammidest, kus teoreetiliselt pole midagi valesti tehtud, kuid mis moonutab tegelikku infot luues visuaalse efekti suurest erinevusest seal, kus seda tegelikult ei ole.

Ebasobiv analüüsimeetod

Kui palju tuntakse erinevaid statistilisi arvnäitajaid ning teste? ... Palju! ... Kahjuks mitte! ... Tavaliselt teatakse ja kasutatakse ikka sedasama tuttavat *keskmist*; vahest ka protsente ja tõsisemal juhul näiteks t-testi või ANOVA't. Ometi hakkab statistikaalase kirjanduse sirvimisel silma, et erinevaid arvnäitajaid ja teste on lõputu hulk. Milleks siis see paljusus? Kas "tavakasutajale" piisab keskmisest ja kõik ülejäänud jäägu nende mängumaaks, kel statistika kui sellise vastu sügavam huvi? ... Mõneti muidugi õige loogika – kui ei tea ega tunne, siis parem ei kasuta! Selle mõtteviisi paratamatuks tulemuseks tundub olema aga see, et kõige tuntumad arvnäitajad ja testid saavad kasutatud ka sellistes olukordades, kus otstarbekam või õigem oleks kasutada hoopis midagi muud. Siit edasi on ainult väga lühike samm saadud tulemuste väärade interpreteerimiseni, mis viitab sellele, et tegelikult ei teata-tunta piisavalt hästi ka nende kõigile tuntud statistikute olemust ning kasutusala.

Statistilise analüüsimeetodi valik sõltub eeskätt muidugi küsimusest, millele me vastust saada tahame (n *Kui suur osa vastanutest ...?*, *Kas kahe grupi tulemused erinevad ...?* *Kas testitulemus on seotud ... ?* jne). Siit edasi sõltub aga meetodi valik veel mitmest asjast. Esiteks sellest, mis tüüpi andmetega on tegu e kas meid huvitav nähtus on mõõdetud arvskaalal või hoopiski kategoriaalsel e sõnaliste väärtustega skaalal. Näiteks küsimustele *Kas tüdrukute ja poiste testitulemused erinevad?* ning *Kas tüdrukute ja poiste muusikaeelistused erinevad?* vastuste saamiseks tuleb ilmselt kasutada erinevaid statistilisi võtteid, sest testitulemused on mõõdetavad arvskaalal, muusikaeelistuste teadaasaamiseks on aga otstarbekam kasutada sõnaliste valikvastustega skaalat. Neile kahele põhilisele tegurile lisandub õige analüüsimeetodi valikul veel terve rida konkreetsemaid tegureid, näiteks: võrreldavate gruppide arv, korraga vaadeldavate tunnuste arv, konkreetse analüüsimeetodi eelduste täidetus jne.

Vahel võib juhtuda ka nii, et tehniliselt tundub meetodi valikuga kõik parimas korras olema ja ka tulemus vastab meie poolt konstrueeritud "statistilisele küsimusele", aga väited, mis me tulemustele rajame, on ikkagi valed. Lihtne näide: kasutada on suhteliselt suur läbilõikeline andmestik, kus muuseas on ka järgmised tunnused:

A. Kui tihti suitsetate (1. iga päev, 2.vahel, kuid mitte iga päev, 3. ei suitseta)

B. Tervislik seisund (1. halb, 2. rahuldav, 3. hea)

Uurija Soss-stat otsustab andmete põhjal leida kinnitust väitele, et suitsetamine kahjustab tervist ehk et suitsetajate tervis on halvem kui mittersuitsetajatel. Lähtuvalt andmete iseloomust ja küsimusepüstitusest valib ta meetodiks risttabeli lootes näidata, et igapäevaste suitsetajate hulgas on enam halva tervisega inimesi. Tulemus on toodud tabelis 1.

| | Kas suitsetate? | | | | | |
|-----------------|-------------------|--------|-------------------|--------|-------------------|--------|
| | iga päev | | vahetevahel | | ei suitseta | |
| | Tervislik seisund | | Tervislik seisund | | Tervislik seisund | |
| | Arv | % | Arv | % | Arv | % |
| halb | 49 | 15.9% | 4 | 6.8% | 111 | 22.0% |
| ei halb ega hea | 140 | 45.3% | 27 | 45.8% | 231 | 45.7% |
| hea | 120 | 38.8% | 28 | 47.5% | 163 | 32.3% |
| Kokku | 309 | 100.0% | 59 | 100.0% | 505 | 100.0% |

Tabel 1. Soss-stati koostatud risttabel suitsetajate ja mittersuitsetajate tervisliku seisundi võrdlemiseks.

Soss-stat on algul küll natukene hämmeldunud, et väide ei leidnud andemete põhjal kinnitust kuid seda suurema rõõmuga teeb ta teatavaks, et tava-arusaam suitsetamise kahjulikkusest ei pea andmete põhjal paika, sest ilmnes tendents, et mittesuitsetavate inimeste tervis on isegi halvem kui suitsetavatel inimestel!

Viimane pool järeldusest on tõepoolest andmete põhjal õige ja tegelikult üsna oodatav ka siis kui suitsetamine tegelikult ikkagi tervist kahjustab (mida ma siiralt edasi usun vaatamata Soss-stati järeldusele)! Viga tekkis juba enne statistilise meetodi valikut ehk Soss-stat ei osanud näha mittevastavust oma “uurimisküsimuse” (*kas suitsetamise kahjustab tervist?*) ning “statistilise küsimuse” (*kas suitsetavate inimeste hulgas on enam halva tervisega inimesi?*) vahel. Mittevastavus tuleneb esiteks andmestiku iseloomust ja teiseks sellest, et seos suitsetamise ja tervisliku seisundi vahel pole ehk sugugi ühesuunaline. Võrreldes suitsetajaid mittesuitsetajatega peaksime küsima *kes?* ja *miks?* antud juhul ühte ja teise gruppi kuuluvad!

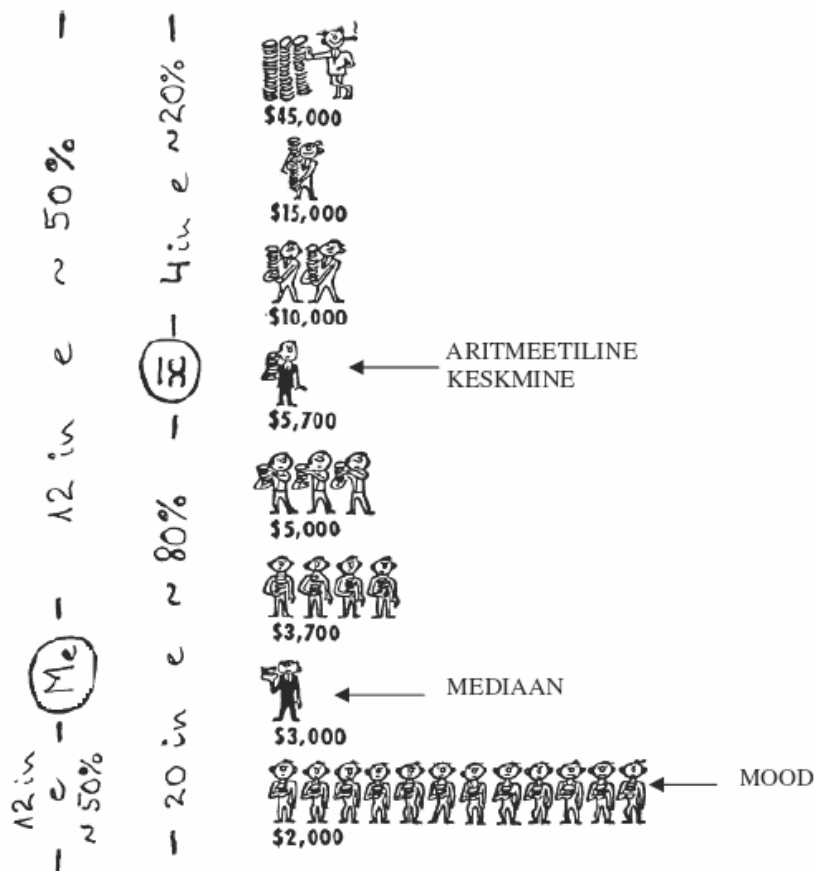
Ilmselt on nii, et suitsetajate hulgas on enam noori inimesi (kelle tervist suitsetamine pole veel jõudnud ära rikkuda) ja mittesuitsetajate hulgas enam vanu (kelle tervis ka ilma suitsetamata on halvem kui noortel) ja ehk ka neid, kes pideva suitsetamise tagajärjel tekkinud tervisehäirete tõttu on suitsetamisest loobunud?! Seega “uurimisküsimusele” vastuse saamiseks peaksime suutma kontrollida vanust ja teadma suitsetamisharjumusi pikema aja jooksul, mitte lähtuma hetke seisust. Kui andmestik seda lisainformatsiooni ei sisalda, siis ei aita ükski statistiline meetod püstitatud “uurimisküsimusele” vastust leida.

Osaline tulemuste kirjeldus e ebapiisav info tulemuste õigeks interpreteerimiseks. Analüüsitulemuste väär interpreteerimine.

Kui proua Mary, kes saab palka 5000 krooni kuus, saab teada, et keskmine palk on 5700 krooni, mida ta mõtleb? Et ta on madalamalt tasustatud töötajate poole hulgas? ... Üsna suure tõenäosusega! ... Tegelikult on tal sellise järelduse tegemiseks vähe põhjust ja alust. Esiteks, on keskmist tendentsi näitavaid arvnäitajaid rohkem kui üks ja igal neist on oma arvutuseeskiri ja tähendus, seega rääkides “keskmisest” ei tea me veel millisest keskmisest jutt käib. Kui oletada, et kasutatud on kõige tuntumat e aritmeetilist keskmist, siis võib aga üsna kindlalt arvata, et Mary järeldus ei ole õige, sest palkade jaotus on tavaliselt küllalt tugeva positiivse asümmeetriaga, millest tulenevalt aritmeetilisest keskmisest kõrgemaid väärtusi on proportsionaalselt tunduvalt vähem kui madalamaid! (vt joonis 4) Seda, kas see antud juhul ka paika peab, saaks teada kui peale keskmise oleks esitatud veel teisi jaotust iseloomustavaid statistikuid.

Keskväertuste ja gruppidevaheliste erinevuste võrdlemise kõrval on statistilise analüüsi eesmärgiks tihti uurida seoseid erinevate nähtuste vahel. Saanud arvutiprogrammilt tulemuse, et tunnuste A' ja B' vaheline seosekordaja on $r=0,18^{**}$ rõõmustava kommentaariga: *** seos on statistiliselt oluline nivool $p<0,01$* , rutatakse pahatihti tegema järeldust laadis: uurimistulemuste põhjal võime väita, et nähtus/omadus A mõjutab märkimisväärselt nähtust/omadust B.

Selline järeldus on problemaatiline mitmest aspektist. Esiteks, viitab see nii seosekordaja tõlgendustavade kui statistilise olulisuse tähenduse mittetundmisele ($|r| < 0,3$ loetakse tavaliselt nõrgaks seoseks ja statistiline olulisus ei muuda seda seost kuidagi suuremaks ja tugevamaks, pigem peegeldab ta antud juhul seda, et valim on olnud suhteliselt suur).

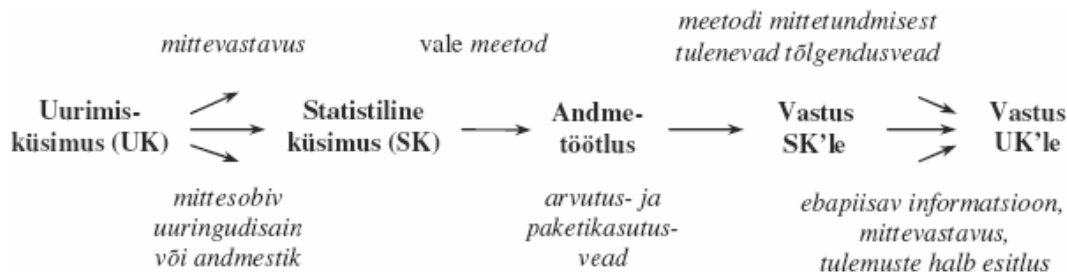


Joonis 4. Erinevad keskmised: keskvärtus, mediaan ja mood tüüpilises palkade jaotuses.

Teiseks, lisainformatsiooni puudumisel on küsitav, kas isegi siis kui seosekordaja oleks näidanud tõepoolest “märkimisväärset” seost tunnuste A’ ja B’ vahel saaksime kinnitada, et see seos on põhjuslik ehk kausaalne. Tugev seosekordaja iseenesest ei räägi midagi kausaalsusest - põhjuslikkus on seose olemasolu korral mõnedel juhtudel tõestatav nõ tavalooigale tuginedes (näiteks kui rahvus või sugu on seotud suhtumisega mingisse küsimusse, siis ilmselt on mõjutavaks teguriks rahvus või sugu ja mitte vastupidi, sest rahvus ja sugu vaevalt et muutuvad sõltuvalt sellest, millise suhtumisega inimene parasjagu on), keerulisematel juhtudel on aga kausaalsuse tõestamiseks vaja vastavat uuringu disaini (eksperimenti), kus teised võimalikud mõjutegurid on kontrollitud.

Kolmas aspekt puudutab seost nähtuse A ja tunnuse A’ ning nähtuse B ja tunnuse B’ vahel. Alati tuleks küsida, kas kasutatavad andmed peegeldavad usaldusväärset kirjeldatavaid nähtusi või mitte, isegi kui see vastavus esmapilgul iseenesest mõistetav tundub.

Kui eelpool räägitu püüda lihtsustatult kokku võtta, siis saaksime järgmise skeemi, millel võimalikud ning ühtlasi vältimist vajavad vead on toodud kaldkirjas:



Joonis 5. Uurimisprotsessi lihtsustatud mudel koos võimalike vigade kirjeldusega.

tulemuste üldistamisel tehtavad vead

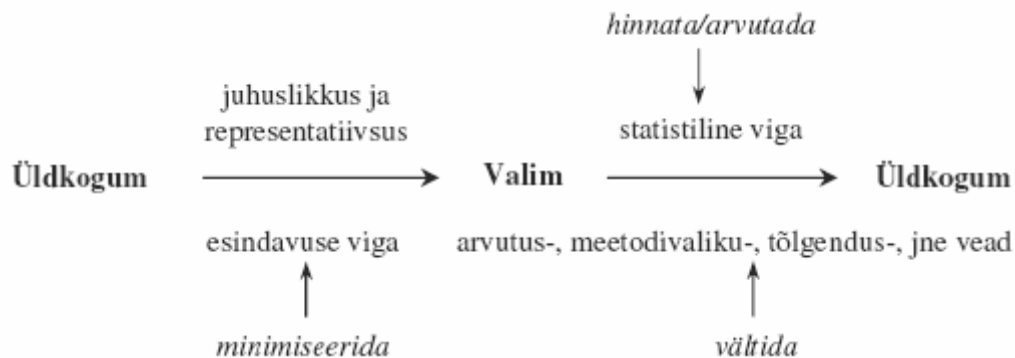
Siiani oleme rääkinud vigadest, mida teha ei tohiks – edasi läheb aga asi keerulisemaks, sest teatud hetkel tulevad statistikas mängu vead, mida me vältida ei saa, mis aga ei tähenda sugugi seda, et me need vead “unustada” võiksime.

Tihti on uuringu eesmärgiks teha järeldusi nii suure hulga inimeste või muude objektide kohta, et kogu selle hulga mõõtmine (st küsitlemine, vaatlemine, testimine jne) ei ole võimalik ja/või mõistlik (näiteks kui soovitakse rääkida kogu eesti elanikkonnast või kõigist Tallinna õpilastest jne). Sellisel juhul valitakse uuringu läbiviimiseks kõigi meid huvitavate objektide (ehk **üldkogumi**) hulgast väiksem alamhulk (ehk valim) ning püütakse hiljem valimi tulemuste põhjal teha üldistavaid järeldusi kogu üldkogumi kohta. Selleks, et niisugused üldistused oleksid usaldusväärsed, tuleb valimi moodustamisel järgida kindlaid reegleid, teisisõnu - valim peaks olema juhuslik ja esinduslik e representatiivne. Juhuslikkus tähendab seda, et kõigil üldkogumi objektidel peab olema võrdne võimalus valimisse valitud saada ja representatiivsus seda, et kõik üldkogumi alamosad peaksid valimis esindatud olema. Representatiivsus on tihedalt seotud valimi suurusega – mida suurem valim, seda lihtsam on tagada tema representatiivsus.

Kui meie valim vastab kirjeldatud nõuetele¹, siis saame statistika abil teha **empüürilisi üldistusi** valimilt üldkogumile, kusjuures vastavate statistiliste meetodite abil saame hinnata üldistamise käigus tekkivat **statistilist viga**. See viga on tingitud sellest, et me ei ole vaadelnud kogu üldkogumit vaid juhuslikku alamosa sellest ning seda viga ei saa vältida, küll aga vähendada – üldjuhul kehtib reegel, et mida suurem valim, seda väiksem statistiline viga! Seega klassikalise (e üldistava e järeldava) statistika² keskseks mõisteks on tõenäosus, s.t. üldistusi tehes ei ole meil kunagi 100% kindlust – statistiliste meetodite kasutamine võimaldab määrata meie eksimise või meie väite kehtimise tõenäosuse. Samas, endiselt tuleb hoiduda kõigist nendest vigadest, millest oli juttu eespool (vt joonis 6).

¹ Praktikas on absoluutset juhuslikkust ja esindavust väga raske tagada, seetõttu on välja töötatud hulk meetodeid, mis aitavad neile nõuetele ligilähedaselt vastavat valimit moodustada.

² ingl: *inferential statistics*



Joonis 6. Vead järeldavas e üldistavas statistikas.

Kui statistiline viga on võrreldes valimis saadud tulemusega (erinevuse või seose või jne -ga) piisavalt väike siis on ka meie eksimise tõenäosus üldistamisel väike ning sel juhul öeldakse, et saadud erinevus või seos või jne on **statistiliselt oluline**. Statistiline ja “eluline” olulisus on aga tihtipeale hoopis erinevad asjad! Statistilist olulisust võib tõlgendada kui luba üldistuste tegemiseks valmilt üldkogumile. Oletame, et meid huvitas erinevus eestlaste ja muulaste hinnangute vahel; meil oli korrektne valim ning saime teada, et valimis ilmnenud erinevus eri rahvusest inimeste hinnangute vahel on statistiliselt oluline. Omamata enam infot võime väita ainult seda, et tulemuste põhjal on ka üldkogumis küllalt suure tõenäosusega olemas erinevus eestlaste ja muust rahvusest inimeste hinnangute vahel. Kas see erinevus on suur ja ka “eluliselt” oluline või vähemalt märkimisväärne, seda tuleb alles uurima hakata!

Tihti arvatakse, et suure(ma)d erinevused, seosed jms on statistiliselt olulised ja väikse(ma)d mitte. Tõepoolest, see ilmne loogika kehtib kuid ühe lisatingimusega: eeldusel, et me räägime ühesuguse suurusega valimitest! Kuna üldistamisel tekkiva vea suurus sõltub valimi suurusest, siis võib juhtuda nii, et **suurte valimite** korral osutuvad statistiliselt olulisteks ka väga väikesed erinevused ja seosed, mis praktika seisukohalt ei ole märkimisväärsed ning vastupidi, **väikeste valimite** korral osutuvad küllalt suured erinevused ja seosed statistiliselt mitteolulisteks, mis ei tähenda, et me peaksime või tohiksime teha järelduse, et tegelikult üldkogumis erinevus või seos puudub - lihtsalt meie valim on nii väike, et me ei saa üldistusi teha! See on koht, kus järelduste tegemisel tihtipeale ekslikult väidetakse, et õnnestus tõestada, et erinevust või seost ei eksisteeri, kuigi õigem oleks väita, et antud uuringutulemuste põhjal ei ole alust väita, et see erinevus või seos üldkogumis olemas on. Kuna tavapäraste statistiliste meetodite abil ei saa ranges mõttes tõestada erinevuse, seose või jne puudumist, siis tuleb ka järeldus statistiliselt mitteolulise tulemuse puhul sõnastada “pehmelt”: ... ei ilmnenud ..., ... ei osutunud oluliseks ... jne.

Kuigi mõned uurijad arvavad, et ilma juhusliku ja esindava ning piisavalt suure valimita ei saa mingisuguseid üldistusi teha, on teised arvamusel, et ka väikeste ja mittejuhuslike valimite põhjal võib mõnes olukorras usaldusväärseid üldistavaid järeldusi teha. Sel juhul on aga tegemist pigem **teoreetiliste üldistustega**, mis ei saa põhineda statistilistel arvutustel vaid lähtuvad uuritud juhtumi(te) tüüpilisusest jms. Nii tuleks sel juhul ka oma üldistusi põhjendada ning hoiduda kasutamast statistilisi meetodeid, mille kasutamise eeldused ei ole täidetud!