TALLINNA ÜLIKOOL

Andmeanalüüsi loengumaterjale:

Andmete esmane töötlemine, analüüsimine ja esitamine

Katrin Niglas
Tallinna Ülikool
informaatika instituut

Sisukord

Sissejuhatus

Andmete kogumise ja analüüsimise viise on mitmeid – tihti räägitakse (ehk pisut liialt üldistades) kvantitatiivsest ja kvalitatiivsest metoodikast. Andmete kogumisest rääkides eelistan sõnapaarile kvantitatiivne – kvalitatiivne kasutada sisult konkreetsemaid märksõnu: struktureeritud ja struktureerimata andmekogumise instrumendid ja/või andmed. Struktureeritud instrumendi tüüpilise näitena võib ette kujutada üht tavapärast ankeeti, kus vastajale on ette antud nii küsimused kui ka võimalikud vastuste variandid, mille hulgast ta vastavalt juhendile sobiva(d) välja peab valima; struktureerimata andmekogumise tüüpilise näitena võib ette kujutada avatud intervjuud, mis sarnaneb vabale vestlusele, kus intervjueerija ei esita konkreetseid lühivastust eeldavaid küsimusi, vaid suunab intervjueeritavat teatud teemadest rääkima, esitab kuuldu põhjal täpsustavaid küsimusi ning julgustab teda oma mõtteid põhjalikult lahti seletama ja põhjendama. Loomulikult võib ette kujutada ka vahepealset varianti, kus vastajale esitatakse kas kirjalikult või suuliselt vastamiseks avatud st ilma vastusevariantideta, kuid küllalt konkreetseid küsimusi, millele eeldatakse vastaja oma tõlgendusest lähtuvat, kuid siiski suhteliselt lühidat vastust. Sellisel juhul võiks rääkida poolstruktureeritud andmekogumise instrumendist.

Käesolevas õppematerjali osas keskendume struktureeritud andmete töötlemiseks, esmaseks analüüsiks ja esitlemiseks sobivatele statistilistele meetoditele.[1] Samas ei ole andmete analüüsimiseks sobivate meetodite valikul määravaks mitte niivõrd see, mis kujul on esialgsed andmed, kuivõrd andmete kohta esitatavate küsimuste olemus. Seega, võib praktikas osutuda vajalikuks struktureerimata andmete selline töötlemine, mille käigus andmetele ”luuakse” sobiv struktuur kodeerimise teel, misjärel saab tekkinud struktureeritud andmeid edasi analüüsida muuhulgas ka statistiliste meetoditega.

Käesolevat materjali täiendab praktiliste näidete ja harjutusülesannetega, kuid ka mõningate õppematerjali raamesse mitte mahtuvate teemade ülevaatliku käsitlusega, loengukursuse aluseks olev slaidiprogramm. Tekstipõhises materjalis toodud diagrammid on kujundatud selliselt, et nad oleksid korrektselt loetavad ka must-valge trüki puhul, mistõttu on välditud erinevate värvide kasutamist ning eelistatud halle toone. Loomulikult võib diagrammide kujundamisel kasutada ka rõõmsamaid värve, mis aitavad sisu emotsionaalsemalt ja seeläbi meeldejäävamalt esitada. Näiteid ja juhiseid diagrammide kujundamise kohta leiad eelpool mainitud slaidiprogrammist.

[1] Andmete analüüsi puudutava osa koostamisel on kasutatud ideid Derek Rowntree raamatust “Statistics Without Tears” (Benguin Books 1991).

1. Mis on statistika ning kuidas oma andmed ja mõtlemine statistilise analüüsi läbiviimiseks ette valmistada?

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. Kuid kas selles on õige süüdistada statistikat?

Paljud statistika õpikud algavad lubadusega, et lugejad ei pea matemaatikast rohkem teadma, kui oskama lihtsalt liita, lahutada, korrutada ja jagada ning asendada toodud valemites tähed õigete numbritega. Sellegi poolest on lugejad, kes pole kõrgema matemaatikaga kokku puutunud, päris kohkunud nähes, et suurem hulk lehtedest on täidetud valemite, võrrandite ja arvutustega. Pahatihti osutuvad arvutuslikud üksikasjad niivõrd aega ja tähelepanu nõudvateks, et lugejad unustavad sootuks üldised ideed, mida need arvutused illustreerima peaks. Sellise olukorra vältimiseks ei pöörata kogu järgnevas käsitluses tähelepanu mitte niivõrd valemitele ühe või teise statistiku arvutamiseks kui püütakse selgitada statistiliste ideede (kontseptsioonide) ja meetodite olemust ning kasutusvaldkondi sõnade, näidete ja jooniste abil.

1.1 Statistiline mõtteviis. Kirjeldav ja üldistav statistika. Üldkogum ja valim.

Statistiline mõtteviis on meile kõigile igapäevasest elust tuttav ja omane. Võtame ühe lihtsa näite: ma ütlen teile, et lähen täna teatrisse kahe kolleegiga, kusjuures üks neist on 190 cm pikk ja teine 165 cm pikk. Millise järelduse te võite kummagi kolleegi soo kohta kõige kindlamini teha, kui teil rohkem mingit informatsiooni ei ole?

* * *

Ma arvan, et te võisite päris veendunult väita, et üks mu kolleegidest, 190 cm pikkune, on mees ja teine, 165 cm pikkune, on naine. Loomulikult võisite te eksida, kuid teil on igapäevasest elust kogemus, et 190 cm pikkuseid naisi on küllalt vähe. Muidugi ei ole te näinud kõiki mehi või kõiki naisi ning te olete märganud, et paljud naised on paljudest meestest pikemad; kuid ometi võite te nähtud meeste ja naiste põhjal küllalt julgelt teha üldistuse ja väita, et üldiselt on mehed pikemad kui naised. Niisiis, enama informatsiooni puudumisel, tundub teile väga tõenäoline, et pikk täiskasvanu on mees ja lühike on naine.

Selliseid lihtsaid näiteid statistilise mõtteviisi kasutamisest võib tuua veel mitmeid. Iga kord, kui te kasutate fraase nagu: “Viimasel ajal olen käinud kinos keskmiselt kaks korda kuus” või “Naised on üldiselt jutukamad kui mehed” või “Mida varem sa kordama hakkad, seda paremini sul eksamil läheb”, teete te statistilise avalduse, kuigi te ei ole sooritanud ühtegi arvutust. Esimeses näites on tehtud kokkuvõte varasematest kogemustest. Teises ja kolmandas näites on aga varasemaid kogemusi üldistatud ning tehtud järeldus tuleviku või vaadeldust laiema sihtrühma kohta.

Tihti, s.h akadeemiliste uuringute läbiviimise raames, on meil aga vaja kirjeldada mingeid nähtusi või nähtuste vahelisi seoseid palju täpsemini, kui me seda teeme igapäevases vestluses. Oma

tähelepanekute põhjal kujunenud oletuste (statistilises sõnastuses HÜPOTEESIDE) kinnitamiseks peame me läbi viima uuringu, mis sisaldab süstemaatilist ANDMETE kogumist antud nähtuse kohta[2], kogutud andmete töötlemist, analüüsimist ning põhjendatud järelduste tegemist.

Lihtsamal juhul, kui meil on olemas meid huvitava sihtgrupi iga liikme kohta andmed ning me saame eeldada, et mõõtmistulemused on täpsed, s.t ei sisalda süstemaatilisi ega ka juhusest tingitud vigu, saame statistiliste meetodite abil oma andmed kokku võtta ja teha uuritud grupi kohta järeldusi, mille paikapidavuse kindluses ei ole vaja kahelda. Nii võime näiteks peale lastevanemate küsitluse tulemuste kokku võtmist väita, et küsitluses osalenud lastevanematest [täpselt] 135 (62%) nõustus sellega, et erivajadustega õpilased peaksid õppima erikoolis, mitte tavakooli klassides.

Samas tuleb osata aru saada, et statistilise maailmavaate keskseks mõisteks on TÕENÄOSUS, s.t statistika ei anna meile alati 100% kindlust andmete põhjal tehtud järeldustes, vaid lubab määrata, kui suur on võimalus ühe või teise sündmuse toimumiseks, meie poolt tehtud järelduse paikapidavuseks, jms. Statistiline mõtteviis on mõistmine, et meie vaatlused (mõõtmised) ei ole alati täiesti täpsed ning, et meie oletus (hüpotees) ning ka andmete põhjal tehtav järeldus võib kehtida näiteks 95-l (või 99-l) juhul 100-st, kuid mitte 100-l juhul 100-st. Näiteks laps, kelle pikkuseks me oleme mõõtnud 162 cm, ei pruugi olla täpselt nii pikk, sest meie mõõteriist ei ole absoluutselt täpne ja me teeme oma tulemustes ümardusi. Seega võib tema pikkus olla kuskil 161,75 cm ja 162,25 cm vahel, kuid mitte täpselt 162 cm. Kui me kasutame olemasolevaid vaatlusandmeid järelduste tegemiseks teiste (mitte mõõdetud) objektide kohta, näiteks juhul, kui me tahame ennustada ühes klassis käivate laste mõõtmisel saadud keskmise pikkuse põhjal teises klassis käivate laste keskmist pikkust, siis on meil võimalus eksida veel palju suurem.

Seetõttu ei saa me oma järeldustes olla alati täiesti täpsed, kuid statistika võimaldab meil määrata võimalike vigade ulatuse ning seda oma järeldustes arvesse võtta. Nii saame vea arvutamiseks õigeid meetodeid kasutades teatud (piisavalt suure) tõenäosusega väita, et lapse pikkus on näiteks vahemikus 162 ± 0,25 cm; ning võime arvutada, et näiteks 99-l juhul 100-st jääb laste keskmine pikkus teises klassis vahemikku 162 ± 3 cm.

Statistika pakub meetodeid väga erinevate küsimuste lahendamiseks ning statistilisi meetodeid võib mitmeti rühmitada, kuid enamuses statistika käsitlustes tõmmatakse selge piir kahe statistika valdkonna vahele:

1. KIRJELDAV STATISTIKA, mis pakub meetodeid (vaatlus)andmetest kokkuvõtete tegemiseks ja olemasolevate andmete kirjeldamiseks ning

2. ÜLDISTAV STATISTIKA, mis kasutab kogutud (vaatlus)andmeid baasina hinnangute ja prognooside tegemiseks (veel) mitte vaadeldud situatsioonide ning kogumite kohta.

Vaatame veelkord neid lauseid igapäevasest elust, mida ma eelpool mainisin. Milliseid nendest on “kirjeldavad” ja millised “üldistavad” kui silmas pidada ülal mainitud tähendust?

- “Viimasel ajal olen käinud kinos keskmiselt kaks korda kuus”
- “Naised on üldiselt jutukamad kui mehed”
- “Mida varem sa kordama hakkad, seda paremini sul eksamil läheb”

* * *

[2] Andmete kogumise temaatika ei mahu antud artikli raamesse, küll aga võite leida mõningaid näpunäiteid ja suunavaid nõuandeid kaasasolevast slaidiprogrammist.

Esimene lause on kirjeldav, teine ja kolmas aga ei piirdu vaid otseselt kogetu kokkuvõtmisega ja teevad üldistuse või ennustuse tuleviku kohta. Selline kahe statistika valdkonna eristamine on tihedalt seotud kahe väga tähtsa mõistega (statistikas): VALIM ja ÜLDKOGUM.

Üldkogumi (ehk populatsiooni) all mõeldakse kõiki juhtumeid või situatsioone, mille kohta uurijad soovivad, et nende poolt saadud järeldused, oletused või prognoosid kehtiksid. Näiteks võivad erinevate valdkondade esindajad tahta teha järeldusi (kõigi) valgete hiirte õppimisvõime kohta; ära arvata erinevatel eksamitel läbipääsevate õpilaste (üld)arvu; ennustada viljasaaki (kõigil) uue väetisega väetatavatel põldudel; uurida (kõigi) Tallinna koolilaste õpimotivatsiooni jne. Nagu te näete, ei mõelda üldkogumi all mitte ainult inimesi, vaid üldkogumi võib moodustada mistahes meid huvitavate sarnaste objektide hulk.

On aga selge, et tegelikus elus ei ole tihti võimalik vaadelda (mõõta, loendada, küsitleda jne.) kõiki meid huvitavaid objekte. Seepärast peab uurija välja valima suhteliselt väikese osa üldkogumist, et selle põhjal teha järeldus kogu üldkogumi kohta. Sellist uurimiseks valitud (suhteliselt väikest) objektide gruppi nimetataksegi VALIMIKS. Näiteks psühholoog, kes uurib valgete hiirte õppimisvõimet, loodab, et saavutatud tulemused ning seega ka järeldused kehtivad kõigi valgete hiirte puhul - mitte ainult praegu olemasolevate, vaid ka veel sündimata hiirte puhul ning ta võib isegi loota, et tema tulemusi võib sedavõrd üldistada, et need selgitaks inimese õppimist.

Seega, paljud uurijad ületavad kättesaadava informatsiooni piiri: nad üldistavad tulemusi valimilt üldkogumile, nähtult ja kogetult mittenähtule ja mittekogetule. Tulles tagasi kirjeldava ja üldistava statistika mõistete juurde, võime öelda, et kirjeldav statistika tegeleb valimi kohta saadud andmete resümeerimise ja kirjeldamisega, üldistava statistika ülesanne on aga järelduste tegemine laiema objektide hulga - üldkogumi – kohta ja/või mõõtmisel tekkiva juhusliku vea hindamine.

Praktikas võib muidugi tulla ette ka olukord, kus uurijat huvitav sihtrühm on suhteliselt väike (või uurimiseks eraldatud ressursid väga suured) ning ta suudab vajalikud andmed koguda (praktiliselt) kõigi rühma liikmete kohta. Sel juhul räägitakse kõiksest uuringust või juhtumianalüüsist, ning eeldades, et andmekogumise meetodid on olnud sellised, mille puhul mõõtmisinstrumendist tingitud juhusliku vea arvestamine ei ole tähtis, võib vajalike järelduste tegemiseks piirduda vaid kirjeldava statistika meetoditega. Kuna sisehindamise puhul on ilmselt valdavalt tegemist just viimase olukorraga, siis piirdub antud peatükk kirjeldava statistika meetodite tutvustamisega.

1.2 Statistiline andmestik. Andmete e tunnuste tüübid.

Vastavalt sellele, mida me uurida tahame, koosneb meie valim kas üksikutest inimestest, koolidest, valgetest hiirtest, kalendrikuudest, mingitest toodetest, kartulipõldudest või millest tahes. Kõiki valimisse kuuluvaid indiviide või üksusi, kelle/mille käest või kohta andmeid kogutakse, nimetatakse statistikas OBJEKTIDEKS. Kõigil ühte valimisse kuuluvatel objektidel on mingid ühised omadused e TUNNUSED, mis meid huvitavad, näiteks: värvus, vanus, hind, kaal, arvamus millegi suhtes, jne[3]. Andmeid koguma asudes, sõnastame meid huvitavate tunnuste kohta küsimusi (nt ”Kui vana te olete?”, ”Kas teie koolis on sisehindamist varem läbi viidud?”) ja viime läbi vajalikud mõõtmised ning eeldame, et andmete

[3] TUNNUSEKS saab statistilise andmetöötluse kontekstis nimetada sellist omadust, mida saab mõõta või mis on juba kokku võetud nii, et iga objekti jaoks on ainult üks vastus ehk üks ühik infot. Sarnases kontekstis räägitakse vahest ka MUUTUJATEST, kuid viimaste puhul võib olla tegemist ka selliste üldiste omaduste või nähtustega, mis on uuringu seisukohast olulised, kuid mille kohta järelduste tegemiseks tuleb koguda igalt objektilt rohkem kui üks ühik infot (nt sotsiaalne staatus, teadmiste tase, verbaalne võimekus, jne).

kogumise käigus saame iga valimi liikme kohta kõik vastused ehk statistika terminoloogiast lähtudes: VÄÄRTUSED. Väärtused on need, mis aitavad meil objekte üksteisest eristada: mõned objektidest on ühte värvi, mõned teist; mõned on naised, teised mehed; mõned on kallimad, teised odavamad, jne.

Oletame näiteks, et teie laps hakkab kooli minema ning teil on vaja välja valida kõige sobivam kool. Millised on need tunnused, mille põhjal te oma valiku teeksite ehk milliseid andmeid te tahaksite erinevate koolide kohta teada, et neist endale sobivaim välja valida?

* * *

Toon mõned küsimused, mis võiksid minu jaoks olulised olla. Teie nimekiri võib olla pikem või lühem, sisaldada osasid toodud küsimustest või kõiki, jne.

- Mis tüüpi kooliga on tegu? (algkool, 9-klassiline kool, 12-klassiline kool)
- Kui kaugel on kool kodust?
- Kuivõrd mugavalt ja turvaliselt on lapsel võimalik kodust kooli jõuda? (koolibuss, ühistranspordi vahend ilma ümber istumiseta, vahetades teel ühistranspordi vahendit, jalutuskäik läbi metsatuka, jne)
- Milline on kooli maine? (väga hea, hea, rahuldav, halb, väga halb)
- Kas on tegu tavalise riigikooliga, erakooliga, või eri(lise)kooliga (nt spordikool)?
- Millised huviringid koolis tegutsevad? (laulukoor, korvpalli trenn, kunstiring, jne)
- Mitu paralleelklassi avatakse?
- Kui suured on selles koolis klassid? (väikesed, keskmised, suured)
- Mis on õpetajate keskmine vanus selles koolis?
- Kas koolis on juurutatud kvaliteedikindlustussüsteem? (jah, ei)

Olles kõne alla tulevate koolide kohta andmed kokku kogunud, tuleb järelduste ja otsuste tegemiseks andmeid analüüsida. Lihtsamal juhul, kui teil on andmeid vähe (antud juhul siis vaid mõne kooli kohta), piisab sellest, et vaatate kõik andmed üle, mõtlete pisut ja jõuategi otsusele st analüüs toimub ilma formaalseid meetodeid kasutamata. Kui aga andmeid on rohkem, siis on mõistlik andmetest ülevaate saamiseks neid mõne sobiva meetodi abil kokku võtma hakata. Nii võib nt peale ankeetküsitluse läbiviimist hakata vastuseid kokku võtma ankeete ükshaaval (korduvalt) läbi lapates ning erinevaid vastuseid loendades. Fragment sellise analüüsi tulemustest võiks välja näha alljärgnevalt:

Lapse toetamine ja järelaitamine õpetaja poolt?

väga rahul	///// ////	*9*
pigem rahul	///// ///// /////	*15*
pigem rahulolematu	///// /	*6*
väga rahulolematu	//	*2*
Arvamus puudub	///	*3*
		Kokku 35 lapsevanemat

Selline tulemuste käsitsi kokku võtmine ja analüüsimine on aga väga aja- ja töömahukas ning jõuab väga harva lihtsast vastuste kokku lugemisest sügavama analüüsini, mille käigus võiks uurida nt ka erinevusi vastajagruppide vahel, arvamuste omavahelist seotust või arvamuste seotust mõnede teiste näitajatega, arvamuste erinevusi eelmiste aastate tulemustega võrreldes, jms. Seetõttu on enne

analüüsima asumist mõistlik andmed sisestada andmetabelisse kasutades selleks mõnd „ruudulise“ töölehega programmi (nt MS Excel, OpenOffice.org Calc, Statistica, SPSS, jne) ning kasutada andmete analüüsimisel arvuti abi. Viimane päästab meid korduvast ja aeganõudvast andmete loendamisest ning võimaldab kiiresti ja mugavalt kasutada samu andmeid uute sisuliste analüüsiküsimuste vastamiseks.

Algandmetest andmetabelit koostades tuleb eelkõige meeles pidada, et õige andmetabel peab olema „askeetlik“ st hästi lihtsa ja alati samasuguse põhistruktuuriga: iga objekt saab endale tabelis ühe rea, iga tunnus omale ühe veeru ning iga väärtus ühe lahtri. Toon kaks näidet andmetabelitest, mis on mõlemad korrektse ülesehitusega, kuigi esimese puhul on tegu kooliõpilaste ning teisel puhul professionaalide poolt koostatud tabeliga.

Sugu	Sünniaeg	Pikkus	Kaal	Keskmine hinne	Hobi	Tähtkuju
N	3.04.1981	160	48	4,31	sport	Jäär
N	14.11.1979	162	53	3,26	Muusika	Skorpion
M	18.02.1980	169	60	3,67	Ujumine	Veevalaja
M	24.01.1980	162	53	4,38	Magamine	Veevalaja
N	23.08.1980	165	55	5	Muusika	Neitsi
N	23.04.1980	169	54	4,31	sport	Sõnn
N	19.08.1980	168	56	4,44		Lõvi
N	21.05.1980	169	57	4,75	Kassid	Sõnn
M	4.08.1980	179	76	3,38	Söömine	Lõvi

ID	A05	C0100	C0200	C0400	C3200	D02	E03	E05
1	6	21	1	4	560	9	4	4
2	6	26	1	2	482	9	3	4
3	6	30	1	2	700	1	3	2
4	6	31	1	2	3000	3	3	3
5	6	33	1	2	2400	3	2	2
6	6	34	1	2	504	9	3	3
7	6	37	1	2	6000	3	2	2
8	6	46	1	2	1000	4	3	3
9	6	47	1	6	3800	1	3	3

Mugava ja paindliku analüüsi tagamiseks tuleb andmetabeli koostamisel arvestada veel mitmete reeglitega, millest olulisemad on järgmised:

* Igale tunnusele/veerule antakse nimi, mis peab olema unikaalne st teistest erinev ning suhteliselt lühike, sest pikkade nimede puhul võtab õigete tunnuste otsimine analüüsi käigus väga palju aega; ei kasutata mitut veergu ühendavaid pealkirju jms!

* Igas lahtris tohib olla ainult üks väärtus e üks ühik infot st mitut vastust ühte lahtrisse sisestada ei tohi! Seega, kui ühe ankeedi küsimuse puhul on vastajal lubatud valida mitu vastusevarianti, annab iga variant andmetabelis eraldi tunnuse/veeru.

* Professionaalid väldivad andmete sisestamist tekstidena ning kasutavad selle asemel vastusevariantide kodeerimist, sest nii hoitakse kokku aega, välditakse sisestusvigu ning hiljem on võimalik andmeid paindlikumalt analüüsida (PS! ilma kodeerimiseeskirja teadmata ei ole sellist andmestikku sisuliselt võimalik analüüsida; professionaalsed arvutiprogrammid lubavad kodeerimiseeskirja sisestada koos andmetega ja oskavad seal olevaid kirjeldusi ka kasutada)

* Ühes veerus tohivad olla ainult üht tüüpi andmed st kui on otsustatud tunnuse sõnaliste väärtuste asemel kasutada arvulisi koode, siis arvude vahele muid sümboleid ei sisestata; puuduva vastuse/ väärtuse jaoks mõeldakse välja sobiv arvuline kood või jäetakse vastav lahter lihtsalt tühjaks.

Kui nüüd uuesti meelde tuletada meie kümmet kooli valikuks olulist küsimust ja kujutleda, et nende andmete põhjal oleks vaja koostada andmetabel, siis, mis oleks tunnuste/veergude arv selles tabelis?

* * *

Ega päris täpset vastust selle küsimusele ei saagi anda, kuna osade küsimuste puhul pole vastusevariantide nimekiri lõplikuna ette antud, aga igal juhul on kindel, et kogu infot ei saa ära mahutada kümnesse veergu, kuna 3. ja 6. küsimuse puhul võib ühe kooliga olla seotud rohkem kui üks vastus, mis viitab vajadusele moodustada andmestikku nende küsimuste jaoks rohkem kui üks tunnus.

Kui nüüd eeldada, et andmestik sai korrektselt koostatud ja andmed sisestatud, siis võiks järgmise sammuna asuda andmeid analüüsima. Selleks on vaja kõige pealt välja mõelda ja enda jaoks selgelt sõnastada küsimused, millele me analüüsi käigus vastuseid saada tahame! Viimane on vajalik selleks, et otsustada, milline meetod on antud olukorras kõige sobivam. Pane tähele, et siin räägime nüüd hoopis teistlaadsetest küsimustest kui olid ankeedis; nt ankeedi küsimus võib olla selline „Kuivõrd olete rahul tunni distsipliiniga?“, analüüsi eeldav küsimus aga „Kui suur osa vastanutest oli tunni distsipliiniga rahul ning kui suur osa mitte?“ või „Kas tüdrukute vanemad oli tunni distsipliini suhtes rahulolematumad kui poiste vanemad?“.

Sageli on aga analüüsi suunava küsimuse täpsest sõnastamisest õige analüüsimeetodi valikuks vähe. Kuna andmed võivad olla väga erineva iseloomuga, siis tuleb meetodi valikul ka seda arvesse võtta; nt kui küsida, „Kas tüdrukute ja poiste testitulemused erinevad?“ või siis „Kas poiste ja tüdrukute hobid erinevad?“, on küsimuse tüüp täpselt sama (meid huvitavad kahe grupi vahelised erinevused), kuid vastuse saamiseks sobiv analüüsimeetod on üsna kindlasti erinev, sest esimesel juhul on tegemist arvuliste andmetega, millest on lihtne arvutada nt keskmine testitulemus poiste jaoks ning võrrelda seda siis tüdrukute keskmise testitulemusega, kuid tüdrukute ja poiste keskmist hobi arvutada pole eriti mõistlik ega mõttekas! Seega, tuleb teisele küsimusele vastuse saamiseks leida mõni teine analüüsi meetod.

Andmete tüüpidest rääkimiseks tuletame meelde ülaltoodud kümme küsimust koolide kohta ning püüame koos mõelda, mille poolest võiks sellistele küsimustele vastustena saadavad andmed omavahel erineda?

* * *

Kas panite tähele, et osad oodatavatest andmetest on esitatavad sõnadena (nt „erakool“, „väga hea“, „jah“, „kunstiring“ jne) ning teised arvudena (nt 5 km, 3 paralleeli, 41 aastat jne)? Selline andmete jagamine sõnadeks ja arvudeks on algatuseks väga hea, sest nii saame juba esimese vihje sobivate meetodite kohta: ilmselt on küsimatagi selge, et kui andmeteks on sõnad, siis ei ole analüüsi käigus mõistlik ega ka lubatud kasutada päris kõiki arvutustel põhinevaid meetodeid, mis mõeldud arvuliste andmete analüüsiks. Kuid mõelda tuleb osata ka vastupidi: mitte iga meetod, mis võib olla andmetest ülevaate saamiseks mugav ja otstarbekas sõnaliste väärtustega andmete puhul, ei pruugi osutuda mõistlikuks arvandmete analüüsimisel.

Päris nii lihtsalt aga ei pääse! Õigeks analüüsimeetodi valikuks tuleb osata teha vahet vähemalt kolmel tunnuste põhitüübil: NIMITUNNUSED, JÄRJESTUSTUNNUSED ja ARVTUNNUSED. Esimesel ja kolmandal neist tüüpidest on aga praktilise andmeanalüüsi seisukohast olulised alamtüübid, mistõttu saame viiese jaotuse, kus tüüpe eristavateks võtmeküsimusteks on see,
- kas vastuseid e väärtusi saab üheselt järjestada või mitte?,
- kas vastustest/väärtustest moodustatud skaalal tekkivad vahemikud on võrdsed või mitte? ning
- kas võimalikke erinevaid vastuseid on vähe või palju?

* **Nimitunnused** (nt rahvus: eestlane, venelane, soomlane, ...)

  NB! Nimitunnusel ei ole väärtused üheselt järjestatavad, järjestustunnusel on!

* **Järjestustunnused** (nt haridustase: algharidus, põhiharidus, keskharidus, ...)

  NB! Järjestustunnusel ei ole väärtuste vahemikud võrdsed, arvtunnusel on!

* **Arvtunnused** (nt vanus: 27 a, 32 a, 51 a, ...)

  Arvtunnused **väheste erinevate väärtustega** (nt laste arv: 0, 1, 2, ...)

  Arvtunnused **paljude erinevate väärtustega** (nt palk: 9264 kr, 10037 kr, 14424 kr, ...)

  NB! Kui nimitunnusel on ainult kaks väärtust, siis väärtuste järjestamise ja vahede võrdsuse probleemi ei teki, mistõttu võib selliseid kahe väärtusega e binaarseid tunnuseid tihti analüüsida arvtunnuste analüüsimiseks sobivate meetoditega!

* **Binaarsed** tunnused (nt sugu: mees, naine)

Vaadake nüüd veelkord üle toodud kümme kooli valikut suunavat küsimust ja kujutledes, et selliseid andmeid oleks teie käsutuses mitte kolme või nelja vaid nt saja kooli kohta, otsustage, mis tüüpi tunnuse moodustavad iga küsimuse põhjal saadavad andmed?

* Mis tüüpi kooliga on tegu? (algkool, 9-klassiline kool, 12-klassiline kool)
* Kui kaugel on kool kodust?
* Kuivõrd mugavalt ja turvaliselt on lapsel võimalik kodust kooli jõuda? (koolibuss, ühistranspordi vahend ilma ümber istumiseta, vahetades teel ühistranspordi vahendit, jalutuskäik läbi metsatuka, jne)

- Milline on kooli maine? (väga hea, hea, rahuldav, halb, väga halb)
- Kas on tegu tavalise riigikooliga, erakooliga, või eri(lise)kooliga (nt spordikool)?
- Millised huviringid koolis tegutsevad? (laulukoor, korvpalli trenn, kunstiring, jne)
- Mitu paralleelklassi avatakse?
- Kui suured on selles koolis klassid? (väikesed, keskmised, suured)
- Mis on õpetajate keskmine vanus selles koolis?
- Kas koolis on juurutatud kvaliteedikindlustussüsteem? (jah – ei)

* * *

Kõige lihtsam on vast alustada sellistest küsimustest, mille vastused on esitatavad arvudena. Küsimus „Mitu paralleelklassi avatakse“ annab arvtunnuse, millel on vähe erinevaid väärtusi, sest sobivad variandid on 1, 2, 3, ja ehk leidub ka koole, kus on 4 või 5 paralleelklassi, kuid rohkemate paralleelidega koole mulle küll ei meenu. Samas kui kokku koguda saja kooli kohta õpetajate keskmine vanus ümardatuna ühe kohani peale koma, siis võib juhtuda, et täpselt sama vastust/väärtust ei ole ühelgi koolil või on kokkulangevad vastused ainult mõnel üksikul juhul[4]. Seega, tegemist on arvtunnusega, millel on palju erinevaid väärtusi. Ka kooli kaugus kodust, kui seda mõõta kilomeetrites ühe komakoha täpsusega, annab ilmselt arvtunnuse, millel on palju erinevaid väärtusi.

Esimene, neljas ja kaheksas küsimus annavad esmapilgul järjestustunnused, kuna kõigi nende puhul on võimalikeks vastusteks sõnad, mis on omavahel üheselt järjestatavad paremuse, suuruse vms alusel (nt „väikses“ klassis on vähem õpilasi kui „keskmises“ ja seal omakorda vähem õpilasi kui „suures“ klassis). Samas võib arutada nii, et kui selliste sõnaliste väärtuste vahed on tajutavad võrdsetena, siis võib vastavat tunnust pidada hoopis arvtunnuseks isegi juhul, kui tema väärtused on sõnad ja mitte arvud. Võtame näiteks tunnuse haridus ja küsime, kas alghariduse ja põhihariduse vahe on sama suur kui põhihariduse ja keskhariduse vahe (ning sealt edasi: kas see on sama suur kui keskhariduse ja kõrghariduse vahe)?

* * *

Ilmselt ei ole vahed eri haridustasemete vahel võrdsed ja seetõttu tuleb haridust ikka järjestustunnuseks pidada. Samas, kui analoogiliselt küsida, kas vahed skaalal väärtustega: <u>väga hea – hea – rahuldav – halb – väga halb</u> on võrdsed, siis osad teoreetilisemalt mõtlevad inimesed ütlevad, et täpset „mõõdulinti“ nende vahede mõõtmiseks ju ei ole ja seetõttu ei saa sellise skaala vahemikke võrdseks pidada, kuid teised jälle vaatavad asjale praktiku pilguga ning väidavad, et vastaja tajub oma vastust valides sellist skaalat võrdsete vahedega skaalana, kuna seal ei ole süstemaatilist väljavenitatust üheski punktis[5] ning sellest lähtuvalt võib neid sõnu käsitleda arvudega sarnaselt st pidada vastavat tunnust arvtunnuseks. Mina kuulun viimaste hulka ja seega peaksin küsimusest „Milline on kooli maine?“, saadud tunnust arvtunnuseks.

Nüüd on järgi veel küsimused, mille põhjal saame koole jagada mingitesse rühmadesse, kuid tekkivad rühmad on sellised, mida ei saa panna <u>üheselt</u> mingisse suuruse, headuse, paremuse vms järjekorda st sellised küsimuse, mille põhjal tekivad nimitunnused. Esmapilgul on sellisteks küsimusteks kolmas,

[4] Pane tähele, et andmed, mida meie kasutame kui algandmeid, võivad olla juba kellegi teise poolt statistilise analüüsi tulemusel saadud koondandmed! Nt selleks, et saada iga kooli kohta õpetajate keskmise vanuse näitajat, tuleb eeldada, et koolidel on olemas andmed iga õpetaja vanuse kohta, mis on keskmise arvutamise valemit kasutades kokku võetud ning sellise koondtulemusena meile meie andmekogumise protsessi käigus kättesaadav.

[5] Selleks peavad toodud variandid andmekogumise instrumendis (nt ankeedis) mitte ainult sisuliselt vaid ka visuaalselt nii esitatud olema, et vahed vastajale võrdsed tunduksid!

viies, kuues ja kümnes. Viienda küsimuse puhul jagatakse koolid nt tavalist õppekava järgivateks riigi- ja erakoolideks, spordikoolideks, muusikakoolideks ja ehk tekivad veel mõned grupid erilisematest koolidest nagu näiteks puuetega lastele mõeldud koolid või alternatiivse didaktilise lähenemisega koolid jms. Iga kool võib kuuluda siin vaid ühte nimetatud gruppidest ja seetõttu saame sellest küsimusest ühe nimitunnuse. Samuti ühe tunnuse saame kümnendast küsimusest, mis puudutab kvaliteedikindlustussüsteemi olemasolu koolis. Kuna nüüd on lubatud vastuseid kaks: jah - ei, siis on tegemist binaarse tunnusega.

Kolmanda ja kuuenda küsimusega on aga lugu pisut keerulisem, sest toodud võimalikest vastustest võib iga kooli kohta kehtida mitu. Ühel tunnusel tohib aga iga objekti (kooli) kohta olla ainult üks vastus e väärtus. Seetõttu tuleb nende küsimuste põhjal esitada alamküsimusi (nt „Kas on võimalik kooli saada koolibussiga?“, „Kas koolis on kunstiring?“, jne) ning teha terve rida binaarseid tunnuseid, millest kõigil on väärtusteks kas jah - ei või on - ei ole.

Andmetüüpidest rääkides on oluline tähelepanu juhtida veel sellele, et kõiki arvtunnuseid on võimalik muuta järjestustunnusteks. Näiteks võime me jagada inimesed pikkuse põhjal gruppidesse: väga pikad, pikad, keskmised, lühikesed ja väga lühikesed. Nii tehes kaotame me aga informatsiooni, ning täpsete algandmete puudumisel me vastupidist teisendust (järjestustunnust arvtunnuseks) teha ei saa. Selline kategoriseerimine on aga tihti vajalik, kui me tahame erinevaid gruppe omavahel võrrelda. Gruppide moodustamist kasutatakse vahel ka selleks, et lihtsustada andmete käsitlemist. Samas andmeid kogudes, tuleks analüüsi seisukohast lähtudes püüda koguda andmeid võimalikult täpselt! See aga, kui täpselt on andmeid võimalik koguda, nii et nende usaldusväärsus oleks piisav, sõltub kontekstist ja olukorrast, kus andmeid kogutakse. Nt võib saada töötajate palga kohta väga täpsed andmed, kui on võimalik kasutada raamatupidamise dokumente, kuid lastevanemate küsitluses ei ole mõistlik küsida inimese täpset sissetulekut kuue viimase kuu lõikes vaid tuleb piirduda etteantud vahemikega ja arvestada vahemike määratlemisel, kas inimene sellises olukorras suudab oma sissetulekuid kokku võtta ja/või on nõus avaldama 100-, 500-, 1000-kroonise või isegi veel väiksema täpsusega.

Tuletame lõpetuseks meelde, miks on vaja oskust oma andmete kohta määratleda, millisega toodud viiest tüübist on iga tunnuse puhul tegemist?

* * *

Nii see, kui kõik teised seni räägitud oskused on vajalikud selleks, et andmeid oleks võimalik mugavalt ja õigete meetoditega analüüsima ning esitlema hakata. Enne analüüsi- ja esitlusmeetodite juurde asumist on hea meelde jätta, et õige meetodi valik sõltub kolmest asjast:

I. Küsimuse tüübist

e mis tüüpi on küsimus, millele tahame analüüsiga vastust saada?
nt Kui **suur osa** vastanutest ...? Kas kolm gruppi **erinevad**? Kas kaks nähtust on **seotud**?

II. Andmete tüübist

e kas analüüsi on kaasatud nimi-, järjestus, arv- või binaarsed tunnused

III. Sihtrühmast

Kui suurt teadlikkust statistiliste meetodite osas võib eeldada?
Milline esitlusviis on selle rühma puhul köitev ja sobilik?

2. Andmete kirjeldamine ehk kuidas saada kogutud andmetest paremat ülevaadet?

2.1 Tabelid ja diagrammid.

Eeldame nüüd, et oleme andmete kogumise ja korrastamise etapid läbinud ja saame alustada andmete analüüsimist. Esimesed küsimused andmete kohta on eeldatavasti üsna lihtsad, sest kõigepealt on vaja andmetest saada üldine ülevaade. Võtame ühe lihtsa näite: kool viis läbi uurimuse, kus üheksandate klasside õpilaste käest küsiti muuhulgas ka seda, millist transpordi liiki nad kooli jõudmiseks kasutavad. Esmased analüüsi eeldavad küsimused võiks olla nt sellised: „Mis on kõige tüüpilisem viis kooli jõudmiseks?“, „Kui suur osa õpilasi tuleb kooli jalgsi?“, „Milliseid transpordi liike üldse kasutatakse ja kui suur on iga transpordivahendit kasutavate õpilaste osakaal?“.

Kõik need küsimused eeldavad vastamist kaht tüüpi küsimustele: kui palju? või kui suur osa? mis eeldab erinevate vastutuste e väärtuste esinemissageduse leidmist e loendamist. Seega, tuleb meil koostada SAGEDUSTABEL, mis võiks antud näite puhul välja näha selline:

Kooli jõudmiseks kasutatavad transpordivahendid

Transpordi liik	Õpilaste arv
Jalgsi	9
Jalgrattaga	3
Mootorrattaga	2
Autoga	6
Trammiga	16
Bussiga	14
Kokku	50

Sellest tabelist saab üsna mugavalt vastused mõnedele ülal välja toodud küsimustele, kuid kas me oskame kiiresti hinnata nende tulemuste põhjal ka jalgsi kooli tulevate laste osakaalu või kui kerge on näha, milliseid transpordi liike kasutatakse rohkem ja milliseid vähem?

* * *

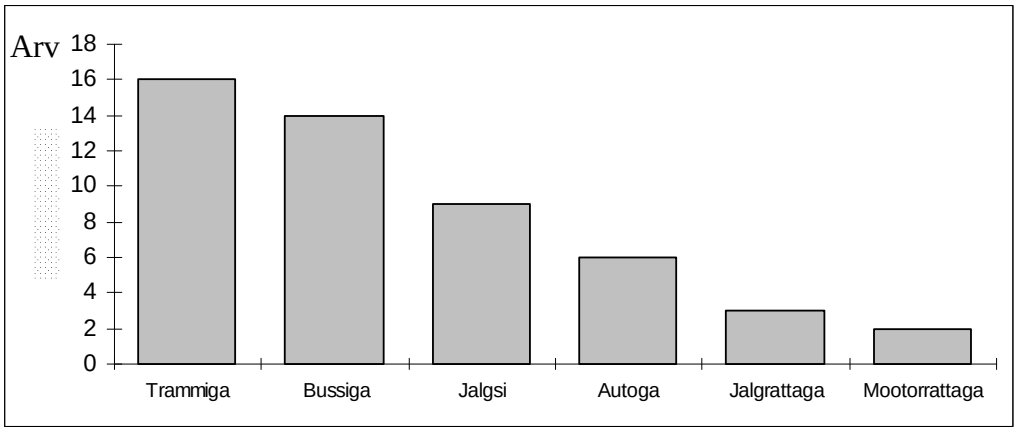
Kuna andmeid on vähe ja osakaalu hindamiseks vajalikud arvutused suhteliselt lihtsad, siis saab muidugi vastused ka nendele küsimustele üsna kiiresti teada, aga kas oleks ehk võimalik andmetest ülevaate saamine lihtsamaks teha? Vaatame alljärgnevat sagedustabelit:

Kooli jõudmiseks kasutatavad transpordivahendid

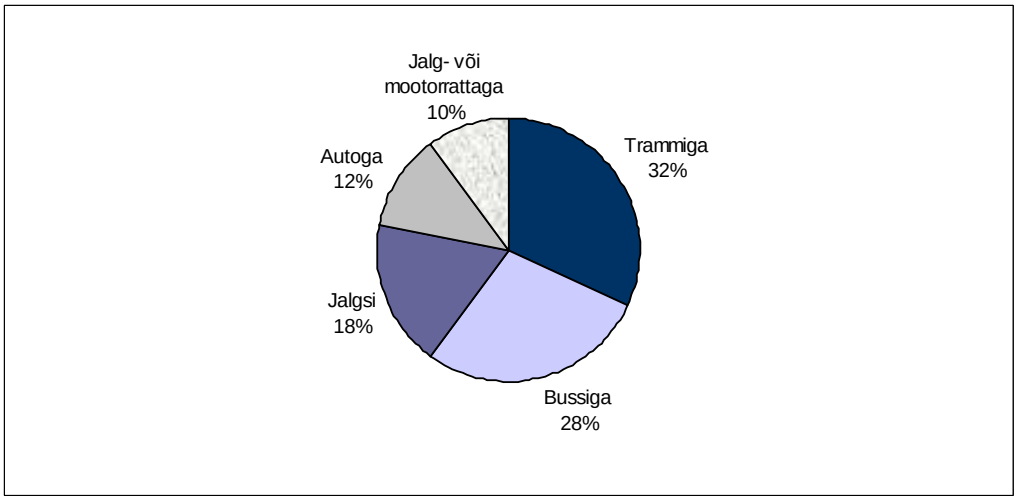
Transpordi liik	Õpilaste Arv	Õpilaste osakaal
Trammiga	16	32 %
Bussiga	14	28 %
Jalgsi	9	18 %
Autoga	6	12 %
Jalgrattaga	3	6 %
Mootorrattaga	2	4 %
Kokku	50	100 %

Tõepoolest, kuna osakaal portsentides on siin selgelt välja toodud ning tabel transpordi liikide esinemissageduse järgi sorteeritud, siis on andmetest ülevaate saamine ning oma küsimustele vastuste leidmine kiirem ja lihtsam kui eelmise tabeli põhjal. Kui nüüd peaks neid tulemusi ka teistele esitlema, siis võiks veelgi sobivaima meetodid üle edasi arutleda ning mõelda, et tabeli asemel võib tulemused esitada ka visuaalselt st diagrammina. Koostame toodud andmetest nt TULPDIAGRAMMI, kus iga tulba kõrgus on proportsionaalne vastavasse kategooriasse kuuluvate õpilaste arvuga:

Kooli jõudmiseks kasutatavad transpordivahendid



Proportsioonide illustreerimiseks kasutatakse ka SEKTORDIAGRAMMI, kus ring on jagatud sektoriteks nii, et iga sektori suurus on proportsionaalne antud kategooria sagedusega.



Kumb diagramm on siis õigem või parem?

* * *

Ühte vastust sellele küsimusele ei olegi. Tihti juhtub nii, et samade andmete analüüsimiseks ja esitlemiseks võib kasutada erinevaid meetodeid, sel juhul tuleb lähtuda sellest, mida eelkõige antud esitlusega rõhutada või esile tuua tahetakse. Antud näite korral oleks vaja arvestada, et tulpdiagramm on ülevaatlikum juhul, kui me eelkõige tahame võrrelda erinevate kategooriate/gruppide sagedusi omavahel, sektordiagramm aga juhul, kui me tahame näidata iga üksiku kategooria/grupi osakaalu tervikus.

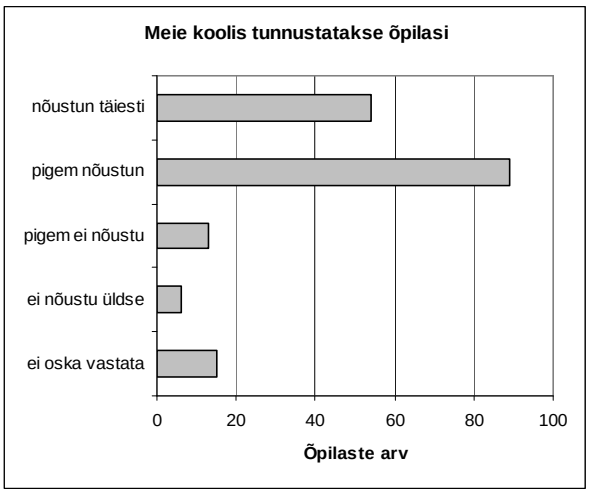
Nägime, et andmete jagunemisest annavad hea ülevaate sagedustabel, tulpdiagramm ja sektordiagramm ning antud näites aitas andmetest kiiremat ülevaadet saada see, kui väärtused tabelis või diagrammil sorteerida sageduste kasvamise või kahanemise järjekorda. Kas need meetodid aga sobivad igat tüüpi andmest ülevaate saamiseks? Mis tüüpi tunnusega oli näites tegu?

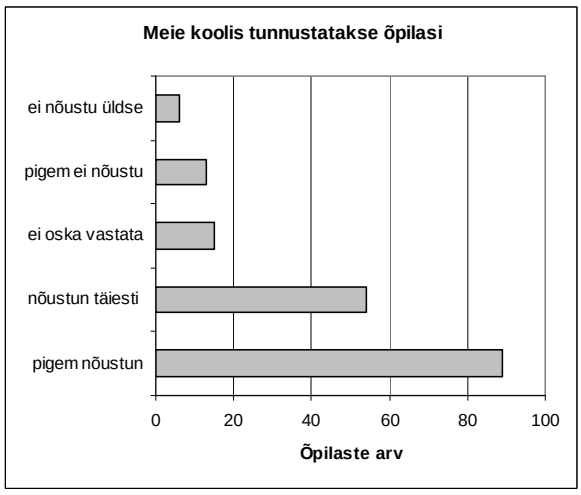
* * *

Tegemist oli nimitunnusega! Sagedustabel, tulpdiagramm ning sektordiagramm sobivad ka järjestustunnuste ja väheste väärtustega arvtunnuste kokku võtmiseks, kuid seda tüüpi tunnuste puhul sageduste järgi sorteerimine head tulemust ei anna, sest siin on väärtused antud sisuliselt loogilises järjekorras, mille „segi ajamine“ teeb tulemist aru saamise raskemaks. Võrdle nt järgmist kahte tabelit või diagrammi:

Meie koolis tunnustatakse õpilasi	Õpilaste arv	Õpilaste osakaal
nõustun täiesti	54	30 %
pigem nõustun	89	50 %
pigem ei nõustu	13	7 %
ei nõustu üldse	6	4 %
ei oska vastata	15	9 %
Kokku	176	100 %

Meie koolis tunnustatakse õpilasi	Õpilaste Arv	Õpilaste osakaal
pigem nõustun	89	50 %
nõustun täiesti	54	30 %
ei oska vastata	15	9 %
pigem ei nõustu	13	7 %
ei nõustu üldse	6	4 %
Kokku	176	100 %





Samas ei sobi ükski ülaltoodud meetoditest ilma vahepeal andmeid teisendamata juhul, kui meil on tegemist arvtunnusega, millel on palju erinevaid väärtusi. Järgmises näites on meil andmed 50 õpilase testitulemuste kohta. Toome tulemused sellises järjekorras, nagu nad testide parandamisel saadi:

50 õpilase testitulemused

89	68	92	74	76	65	77	83	75	87
85	64	79	77	96	80	70	85	80	80
82	81	86	71	90	87	71	72	62	78
77	90	83	81	73	80	78	81	81	75
82	88	79	79	94	82	66	78	74	72

Ma arvan, et te ei vaidle mulle vastu, kui ma ütlen, et sellisel kujul on nendest numbritest peaaegu võimatu midagi välja lugeda. Kas te saate ülevaate õpilaste testitulemustest? Kui kerge on leida kõige kõrgemat ja kõige madalamat testitulemust? Kas testitulemused on jagunenud ühtlaselt minimaalse ja maksimaalse väärtuse vahel või on mõned testitulemused tihedamini esinevad kui teised?

* * *

Ilmselt on siingi vaja andmetest ülevaate saamiseks kokkuvõtteid teha. Kui aga arvutilt selliste andmete puhul tellida sagedustabel, siis tuleb see pea terve lehekülje pikkune ning sektordiagramm meenutab kirjut lõngakera, sest algoritmi kohaselt kantakse tabelisse ridadeks või diagrammile sektoriteks/tulpadeks ühe kaupa kõik erinevad testitulemused, mida on ju väga palju! Seetõttu tuleb vaatlusandmed enne tabelisse või diagrammile kandmist grupeerida. Näiteks võime me küsida mitu testitulemust on vahemikus 60-st 64-ni, mitu 65-st 69-ni, mitu 70-st 74-ni jne. Kui me oma andmeid niimoodi grupeerime[6], saame järgmise tabeli:

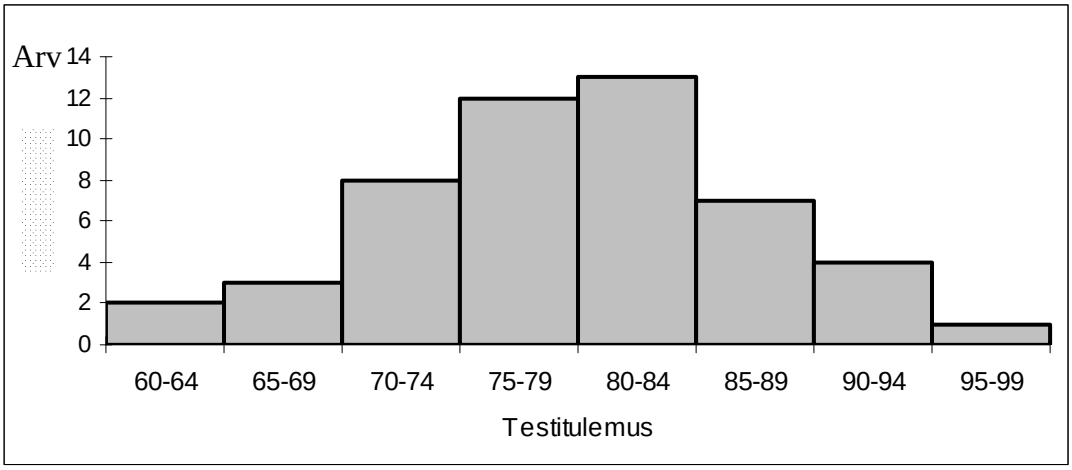
Testitulemus	Õpilaste arv (sagedus)	Õpilaste osakaal (protsent)
60-64	2	4
65-69	3	6
70-74	8	16
75-79	12	24
80-84	13	26
85-89	7	14
90-94	4	8
95-99	1	2
Kokku	50	100

Sellest tabelist on jaotuse üldine kuju selgelt näha - meie näites “kuhjuvad” testitulemused jaotuse keskel, madalaid ja kõrgeid tulemusi on aga vähestel õpilastel. Kuid tuleb tähele panna, et selline grupeerimine toob endaga paratamatult kaasa informatsiooni kao. Jaotuse üldise kuju selgitamisel „tuuakse ohvriks“ üksikud väärtused ja seepärast ei saa niisuguse tabeli põhjal vastata ka kõigile küsimustele, mis meil nende andmete kohta tekkida võivad. Seetõttu tuleb kaaluda (ka) teiste analüüsimeetodite kasutamist, millest tuleb juttu järgmises alalõigus.

Ülaltoodud tabeli graafiliseks esituseks on HISTOGRAMM. Histogramm on tulpdiagrammi spetsiifiline alamliik, kus telgede tähendus on vastupidiselt tulpdiagrammile, mille abil võib esitada väga erinevaid arvandmeid, üheselt määratud. Kui tulpdiagrammil on sellel teljel, millele tulbad toetuvad, tunnuse üksikute väärtuste poolt määratud grupid, siis histogrammil on samal teljel arvutunnuse väärtustest moodustatud vahemikud. Kuna seal, kus lõppeb eelmine vahemik, algab kohe järgmine, siis on histogrammil õige tulbad asetada vahetult üksteise kõrvale, kuna aga tulpdiagrammil võrreldavate gruppide vahel selline pidevus puudub, siis on mõistlik seal jätta iga tulba vahele pisut tühja ruumi. Tulba kõrgust kirjeldava telje tähendus tulpdiagrammil võib väljendada sisuliselt mistahes ühikuid või arvandmeid – historgrammil väljendab tulpade kõrgus alati antud vahemiku sagedust e seda, mitu tulemust (või kui suur osa tulemustest) antud vahemikku jäi.

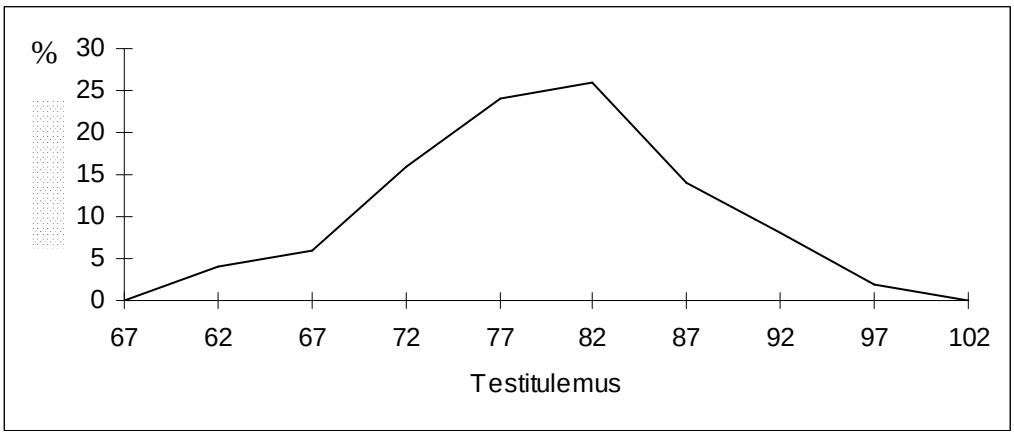
[6] See, kuidas sellist grupeerimist praktikas läbi viia, sõltub kasutatavast arvutiprogrammist. Reeglina tuleb statistikapakette kasutades moodustada olemasoleva põhjal uus tunnus, milles on iga objekti/õpilase jaoks väärtusena kirjas, mitmendasse punktivahemikku tema tulemus jäi.

50 õpilase testitulemuste jaotus (HISTOGRAMM).



Sagedusjaotust võib esitada ka teistsuguse joonise abil. Kanname iga vahemiku keskpunkti kohale punkti - punkti kõrguse määrab jällegi iga vahemiku sagedus või osakaal - ning ühendame saadud punktid murdjoonega. Niisugust joonist nimetatakse (empiiriliseks) JAOTUSKÕVERAKS. Pane tähele, et sama moodi nagu histogramm ei ole tavaline tulpdiagramm, ei ole ka see joonis tavaline joondiagramm, mille puhul alumisel teljel oleksid üksikväärtused ja mitte väärtuste vahemikud.

50 õpilase testitulemuste jaotus (JAOTUSKÕVER).



Sagedustabelite abil saab vastuseid ka pisut keerulisematele küsimustele, kui me seni oleme vaadanud. Mõni aeg tagasi viidi läbi uuring, kus vaadeldi Eesti õpetajate hoiakuid ja tõekspidamisi ning muuhulgas paluti vastajatel 5-palli skaalal hinnata kuivõrd meeldivaks nad peavad autoritaarset juhtimisstiili. End uurija rolli mõeldes võib nüüd püstitada mitmeid analüüsi suunavaid küsimusi, kuid vaatleme ühte neist: „Kas ja kuidas erinevad eesti ja vene rahvusest õpetajate arvamused autoritaarse juhtimisstiili kohta?“. Analüüsimeetodi valikuks peab läbi mõtlema, kas sellist tüüpi andmete puhul sagedustabel sobib ülevaate saamiseks eesti ja vene õpetajate arvamustest?

* * *

Sobib küll! Tegelikult on sagedustabeli kasutamise juures ainult üks piirang – analüüsitavatel tunnustel ei tohi olla liiga palju väärtusi, mis tabeli „välja venitavad“ ning ebaülevaatlikuks teevad. Meie näites on ühel

tunnusel kaks väärtust (eesti ja vene) ning teisel viis väärtust, kuna oli öeldud, et kasutati 5-palli skaalat. Eeldades, et uuringu andmestik oli korralikult arvutisse sisestatud, jääb meil üle vaid paar hiireliigutust teha ja alljärgnev võrdlev sagedustabel annab meie küsimustele vastused. Millise vastuse te alljärgnevast tabelist välja loete?

Autoritaarne juhtimisstiil	eesti õpetajad		vene õpetajad	
	Arv	Osakaal	Arv	Osakaal
täiesti meeldiv	42	10,3%	81	28,0%
Põhiliselt meeldiv	80	19,7%	74	25,6%
osalt meeldiv, osalt ebameeldiv	147	36,1%	97	33,6%
Põhiliselt ebameeldiv	80	19,7%	20	6,9%
täiesti ebameeldiv	58	14,3%	17	5,9%
Kokku	407	100,0%	289	100,0%

* * *

Erinevused eesti ja vene õpetajate arvamuste vahel on märgatavad, kusjuures meeldivamaks peavad autoritaarset juhtimisstiili vene õpetajad; eesti õpetajate seas leidub üsna võrdselt nii neid, kellel selline juhtimine meeldib, kui ka neid, kes seda ebameeldivaks peavad; ligikaudu kolmandik nii eesti kui vene õpetajatest leiab, et autoritaarsel juhtimisstiilil on nii meeldivaid kui ebameeldivaid külgi.

Loodan, et toodud näidete põhjal tekkis teil tunne, et tabelid ja diagrammid on lihtne ja kõigile arusaadav viis oma andmetest esmaseid kokkuvõtteid teha ning tulemusi ka teistele esitleda. Nii see peakski olema, kuid pahatihti juhtub praktikas nii, et lihtsad põhitõed kipuvad meelest minema või siis on korraga nii paljudele pisidetailidele vaja tähelepanu pöörata, et miskit ikka kahe silma vahele jääb ja sassi läheb. Ka see võib probleemiks saada, et teile endale on esitletav sisu nii selge, et te ei oska ennast asjasse pühendamatu vaataja rolli mõelda ning näha, et teie joonise või tabeli kujundus pigem töötab selle vastu, kui et aitab vaatajal sisust lõpuni ja õigesti aru saada. Seetõttu soovitan läbi vaata ja mis veelgi tähtsam – läbi mõelda! – materjaliga kaasas olevas slaidiprogrammis toodud näited ja kommentaarid ning kokkuvõtvad juhised selle kohta, mida tabelite ja diagrammide kujundamise juures eelkõige silmas pidada![7]

2.2 Keskmist tendentsi ja hajuvust väljendavad arvnäitajad.

Nagu eelmises alalõigus mainitud, on mõnes olukorras andmete analüüsimiseks sagedustabelite kõrval või koguni nende asemel sobilikum kasutada arvnäitajaid. Eriti kerkib see vajadus esile, kui tegeleme arvutunnustega, millel on palju erinevaid väärtusi, nagu näiteks andmed palkade või testitulemuste kohta. Suurem osa arvnäitajatest ongi mõeldud kasutamiseks arvutunnuste korral, kuid leidub ka selliseid, mida saab kasutada järjestustunnuste või koguni nimitunnuste puhul.

Vaatame uuesti näidet, kus meil olid andmeteks 50 õpilase testitulemused. Jätame seekord andmete koondamise vahemikesse tegemata ja vaatleme tulemusi üksikväärtustena. Parema ülevaate saamiseks JAOTUSEST e sellest, milliseid tulemusi/väärtusi kui palju on, võime tulemused järjestada kasvamise või kahanemise järjekorda saades niimoodi VARIATSIOONIREA.

[7] Vastava sisuga slaidiprogrammi leiad ka antud peatüki autori kodulehelt www.tlu.ee/~katrin/ õppematerjalide alalõigust.

50 õpilase testitulemused (VARIATSIOONIRIDA)

62	64	65	66	68	70	71	71	72	72
73	74	74	75	75	76	77	77	77	78
78	78	79	79	**79**	**80**	80	80	80	81
81	81	81	82	82	82	83	83	85	85
86	87	87	88	89	90	90	92	94	**96**

Nüüd on meil lihtne saada ülevaade sellest, mis oli kõige madalam ja kõige kõrgem tulemus st leida minimaalne ja maksimaalne väärtus: vastavalt 62 ja 96 palli ning rääkida sellest, kui suured olid tulemuste omavahelised erinevused läbi kõige suurema erinevuse st jaotuse ULATUSE arvutamise. Selleks tuleb leida maksimaalse ja minimaalse väärtuse vahe: meil 96 miinus 62 annab ulatuseks 34 palli.

Sellisest kasvavas järjekorras antud vaatlustulemuste reast on kerge leida ka jaotuse keskel paiknevat väärtust ehk MEDIAANI. Mediaan on selline väärtus, mis jagab vaatlustulemused kahte ossa nii, et pooled vaatlustulemused on mediaanist väiksemad ja pooled suuremad. Seega, kui meil on teada seitsme õpetaja kohta nende keskmine raamatukogus töötamise aeg nädalas (tundides):

0 2 3 4 6 6 10

siis saame öelda, et mediaan on 4 (tundi nädalas).

Kui meil on aga paaris arv vaatlustulemusi, siis ei saa me nende hulgast leida ühte, millest oleks võrdne arv väiksemaid ja suuremaid väärtusi. Seepärast leitakse sel juhul väärtus, mis asub täpselt kahe variatsioonireas keskel asuva väärtuse vahel. Meie näites õpilaste testitulemuste kohta on 25-es väärtus 79 ning 26-es 80. Et leida täpselt nende vahel paiknevat väärtust, tuleb need väärtused kokku liita ning jagada kahega: $\frac{79+80}{2} = 79{,}5$. Seega mediaaniks on 79,5 palli. Viimasest arvnäitajast saame teha nüüd omakorda sisulise tõlgenduse ja öelda, et poolte õpilaste testitulemus jäi alla 79,5 palli ja poolte õpilastel oli see üle 79,5 palli.

Mediaan on üks statistikas kasutatavaid keskmist tendentsi väljendavaid suurusi. Kuid märksa sagedamini kasutatakse ARITMEETILIST KESKMIST, mida tavaliselt kutsutakse lihtsalt keskmiseks või siis keskväärtuseks. Aritmeetilise keskmise leidmiseks tuleb kõik vaatlustulemused kokku liita ning saadud summa jagada vaatlustulemuste arvuga. Leiame nüüd õpetajate raamatukogus töötamise aja aritmeetilise keskmise:

$$\bar{x} = \frac{0+2+3+4+6+6+10}{7} = \frac{31}{7} \approx 4{,}4 \text{ tundi nädalas.}$$

Kui meil on aga teada, et algandmetena kasutatud arvud ei olnud täpsed vaid ümardatud või hinnangulised (st õpetajad ei pruugi raamatukogus töötada täpselt 2 või 6 tundi vaid ligikaudu toodud arv tunde) siis peame ka arvnäitaja põhjal järeldust tehes jääma algandemete täpsuse tasemele ja ütlema, et keskmiselt töötavad õpetajad raamatukogus 4 kuni 5 tundi nädalas.

Kui nüüd kokku liita ka meie 50 õpilase testitulemused ning jagada summa 50-ga, siis saame keskmiseks testitulemuseks 79,1 punkti. Võrreldes kahte erinevat keskmist tendentsi väljendavat suurust: mediaani ja aritmeetilist keskmist, näeme, et nad on natuke erinevad. Tuleme nende võrdlemise juurde mõne aja pärast tagasi ja vaatame siinkohal veel üht keskmist, mis on statistikas küllalt laialt kasutusel.

Kui leiame sellise väärtuse, mida esines teiste väärtuste seas kõige rohkem, saame teada MOODI. Meie testitulemuste näites on kaks moodi, sest nii tulemus 80 kui 81 punkti on saadud nelja õpilase poolt ning pole ühtki sellist tulemust, mis oleks esinenud rohkem kui neljal õpilasel st need testitulemused on kõige "moodsamad" ehk kõige sagedamini esinevad. Samas, kui me endalt küsime, kas neli tulemust 50-st on nii suur osa, et sellest eraldi rääkida ja selle kohta suuri sisulisi järeldusi teha, siis on vastuseks ilmselt „ei". Seega, kuigi arvtunnuste puhul on moodi leidmine tehniliselt lubatud, ei ole see sisulisest küljest tihti otstarbekas. Moodi kasutatakse kõige rohkem nimitunnuste jaotuse iseloomustamiseks, kuigi ta „töötab" hästi kõigi tunnuste puhul, millel on vähe erinevaid väärtusi. Oletame näiteks, et 50-st küsitletud õpetajast 22 olid abielus, 17 vallalised ning 11 lahutatud. Mood on siin "abielus" ehk kõige enam oli õpetajate seas abielus inimesi. Aga kas ka selliselt sõnastatud järeldus oleks moodi põhjal õige: „Suurem osa õpetajatest olid abielus"?

*** * ***

Tõepoolest ei oleks, sest 22 õpetajat 50-st on ju all poole!

Ehk olete märganud, et tihti on testitulemustel jm mõõtmise ehk andmekogumise käigus saadud väärtustel kalduvus koonduda mingi ulatuse keskosas paikneva väärtuse ümber st, et mõõtmisel saame me palju rohkem keskmise suurusega tulemusi kui väikeseid või suuri. Sellist vaatlustulemuste koondumise tendentsi nimetatakse keskmiseks tendentsiks. Eelnevas tutvusime me juba kolme arvnäitajaga, mis seda tendentsi iseloomustavad. Need kolm keskmist on: mood, mediaan ja aritmeetiline keskmine ehk keskväärtus. Millist neist kolmest kasutada, sõltub peamiselt iseloomustatava tunnuse tüübist. Millist keskmist saab kasutada järgmise andmetüübi puhul?

Transpordi liik	Õpilaste Arv	Õpilaste osakaal
Trammiga	16	32 %
Bussiga	14	28 %
Jalgsi	9	18 %
Autoga	6	12 %
Jalgrattaga	3	6 %
Mootorrattaga	2	4 %
Kokku	50	100 %

*** * ***

Sellise nimitunnuse puhul saab keskmistest kasutada ainult moodi. Kõige populaarsem transpordivahend kooli jõudmiseks ehk mood on siin "tramm".

Kui nimitunnuste puhul saab keskmist tendentsi väljendada ainult moodi abil, siis arvandmete puhul on võimalik leida kõik kolm erinevat keskmist. Kõige enamkasutatav keskmist tendentsi väljendav suurus on keskväärtus, sest ta on teoreetilises plaanis kõige stabiilsem s.t võttes ühest üldkogumist erinevaid valimeid muutub keskväärtus mediaani ja moodiga võrreldes kõige vähem. Siit järeldus, et valimi keskväärtus iseloomustab üldkogumit paremini kui mediaan või mood.

Sellegi poolest on situatsioone, kus keskmist tendentsi on õigem iseloomustada mediaani abil või anda valimi kirjeldamiseks nii keskväärtus kui mediaan. Vaadake kahte alljärgnevat jaotust. Mõlemas on toodud viie inimese kuupalgad:

I	8000 kr.	10 000 kr.	14 000 kr.	17 000 kr.	19 000 kr.
II	7000 kr.	11 000 kr.	13 000 kr.	16 000 kr.	39 000 kr.

Mediaanid kahes grupis on küllalt sarnased: I –> 14 000 kr., II –> 13 000 kr. Arvutades aga välja keskväärtused saame, et keskväärtus esimeses grupis on 13 600 krooni, mis on mediaaniga küllalt sarnane, kuid teises grupis on keskväärtus 17 200 krooni, millest on kõik peale ühe väärtuse madalamad.

Esimese grupi puhul saame me nii mediaani kui keskväärtuse abil õige ettekujutuse grupi liikmete keskmisest palgast. Kuid kumb keskmistest annab parema ettekujutuse tüüpilisest palgast teises grupis?

* * *

Teises grupis tuleks keskmist tendentsi väljendava suurusena (keskväärtusele lisaks) kasutada mediaani, sest keskväärtus on tugevalt mõjutatud ühest ebatüüpilisest, teistest väga erinevast väärtusest, mediaani aga sellised ekstreemsed väärtused ei mõjuta.

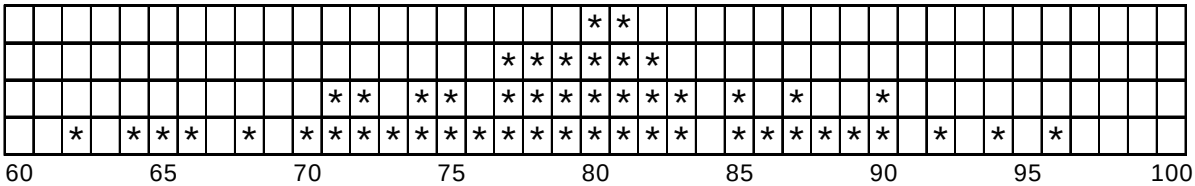
Kuigi keskmised on kõige tuntumad ja enamkasutatavad arvnäitajad, ei anna ainult keskmise teadmine meile andmete kohta täit pilti. Jaotuse iseloomustamisel on väga tähtis tähelepanu pöörata ka sellele, kuivõrd erinevad või sarnased on tulemused/väärtused omavahel. Vaatame ühte näidet, kus lastevanematel paluti 7-palli süsteemis hinnata kuivõrd tähtsaks nad peavad seda, et kool lastes arendaks järgmisi väärtusi:

1)	Kohuse- ja vastutustunne	(viie lapsevanema vastused: 3 4 4 4 5)
2)	Aktiivsus, ettevõtlikus	(viie lapsevanema vastused: 1 2 3 7 7)

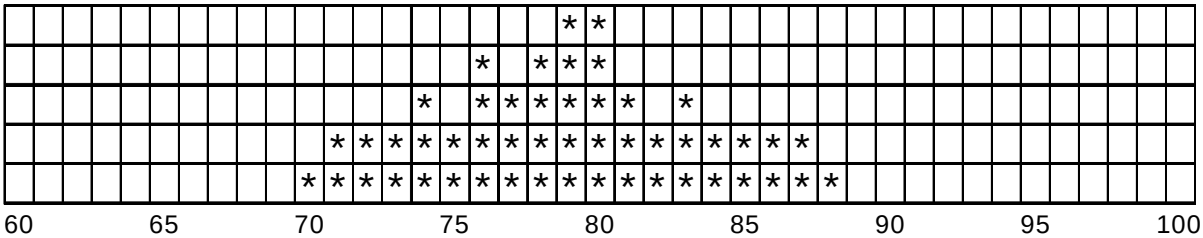
Mõlemal juhul on keskmiseks tähtsuse hinnanguks 4 palli, kuid ometi näeme selgelt, et lastevanemate arvamused nende kahe aspekti arendamise tähtsuse osas ei ole täpselt ühesugused: kohuse- ja vastutustunde arendamise osas on lapsevanemad olnud suhteliselt üksmeelselt arvamusel, et see on keskmise tähtsusega, kuid aktiivsuse ja ettevõtlikkuse arendamist on osad lastevanematest pidanud väga tähtsaks, teised jälle üldse mitte tähtsaks st vastajate arvamused on olnud väga erinevad. Sellist väärtuste omavahelise erinevuse määra nimetatakse statistikas HAJUVUSEKS. Hajuvus ongi keskmise kõrval teine oluline jaotust iseloomustav suurus.

Et hajuvuse mõistest paremat ettekujutust saada, võrrelge kahte järgnevat punkt-diagrammi, kus on kujutatud kahe erineva õpilasterühma testitulemused:

50 õpilase testitulemused - GRUPP A



50 õpilase testitulemused - GRUPP B



Mis on teie arvates kõige suurem erinevus nende kahe jaotuse vahel? Kas te oskate öelda, milline juba vaadeldud arvnäitajatest aitab seda erinevust kirjeldada?

* * *

Diagrammidele peale vaadates võime kohe näha, et esimene jaotus on rohkem „välja venitatud“ st testitulemused grupis A on rohkem hajunud kui grupis B. Jaotuse HAJUVUST ehk VARIATIIVSUST saame kõige lihtsamini väljendada arvutades jaotuse ulatuse. Meie näites:

grupis A on ulatus = 96 - 62 = 34 punkti

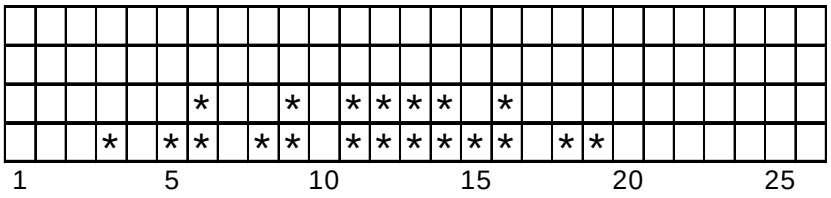
grupis B on ulatus = 88 - 70 = 18 punkti

Seega, saaksime ulatuse põhjal ka siis, kui meil andmetest diagrammi tehtud ei ole, teha järelduse, et grupis B on tulemuste omavahelised erinevused e hajuvus palju väiksem kui grupis A.

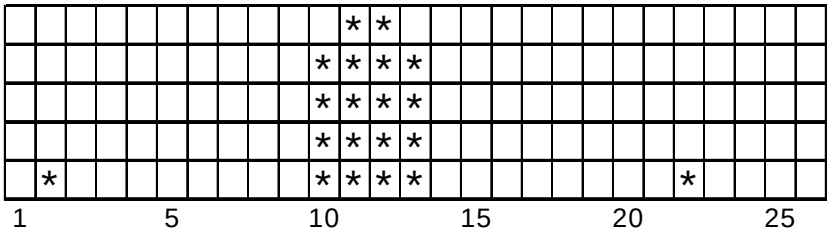
Ulatus on kõige üldisem ja lihtsamini leitav hajuvuse näitaja, kuid tema suur puudus on selles, et ta sõltub ainult jaotuse kahest kõige äärmisest väärtusest, mis võivad aga mingil põhjusel olla teistest väga erinevad nn ekstreemsed väärtused (tuletage meelde näidet palkadest!). Seepärast on selle näitaja usaldatavus grupi kui terviku iseloomustamisel väike ning teda kasutatakse vaid jaotusest kõige üldisema pildi saamiseks.

Vaatame veel kahte punktdiagrammi, kus on kujutatud kahe üliõpilaste grupi (mõlemas 20 õpilast) testitulemused:

Grupp I



Grupp II

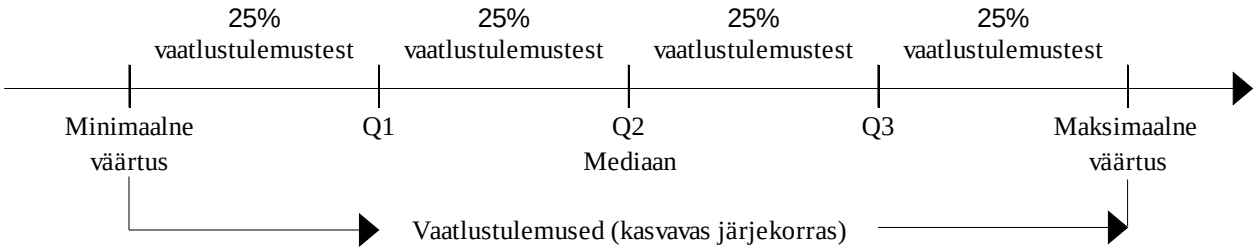


Kumb toodud jaotustest on teie arvates suurema hajuvusega? Kas ka ulatus selles jaotuses on suurem?

* * *

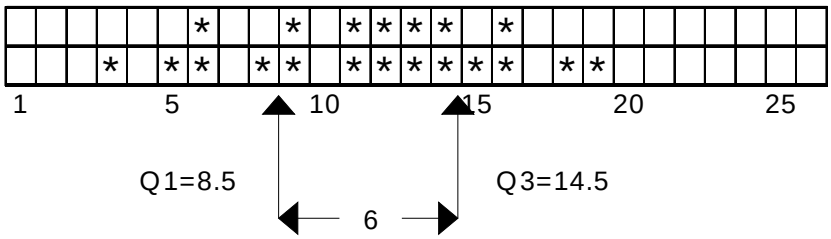
Üldiselt tundub, et esimese grupi tulemuste hajuvus on suurem, sest selles grupis on saadud 13 erinevat punktisummat, kusjuures teises grupis, kui välja jätta kaks ekstreemset väärtust, on saadud punktid jaotunud väga ühtlaselt ainult nelja erineva väärtuse vahel. Siiski on teise grupi ulatus tänu ekstreemsetele väärtustele (ühele väga heale ja ühele väga halvale tulemusele) suurem kui esimese grupi oma.

Üks võimalus leida “paremat” hajuvuse näitajat on vaadelda mingit väiksemat jaotuse keskpunkti ümber asuvat väärtuste piirkonda, mis võimaldab teistest tugevalt erinevate väärtuste mõju kõrvaldada. Sellise piirkonna moodustamisel on meile abiks KVARTIILID. Kui mediaan jagab meie vaatlustulemused kahte võrdsesse ossa, siis kvartiilid võimaldavad need jagada nelja võrdsesse ossa nii, et igasse ossa jääb 25% tulemustest:



Seega on kokku kolm kvartiili, kusjuures teine kvartiil on võrdne mediaaniga. Esimest kvartiili nimetatakse ka alumiseks kvartiiliks ning kolmandat ülemiseks kvartiiliks. Jaotuse hajuvuse kirjeldamiseks kasutatakse kvartiilide vahet: $Q_3 - Q_1$.

Meie näites on mõlemas grupis 20 õpilast, seega esimene kvartiil lõikab ära 20 / 4 = 5 väiksemat väärtust ning kolmas kvartiil 5 suuremat väärtust. Esimeses jaotuses on viies väärtus 8 ning kuues 9. Seega Q_1 = 8,5. Samuti, kuna viieteistkümnes väärtus on 14 ja kuueteistkümnes on 15, siis Q_3 = 14,5.



Kvartiilide vahe on aga 14,5 - 8,5 = 6 palli.

Leidke nüüd kvartiilide vahe teises grupis ning võrrelge saadud tulemusi omavahel?

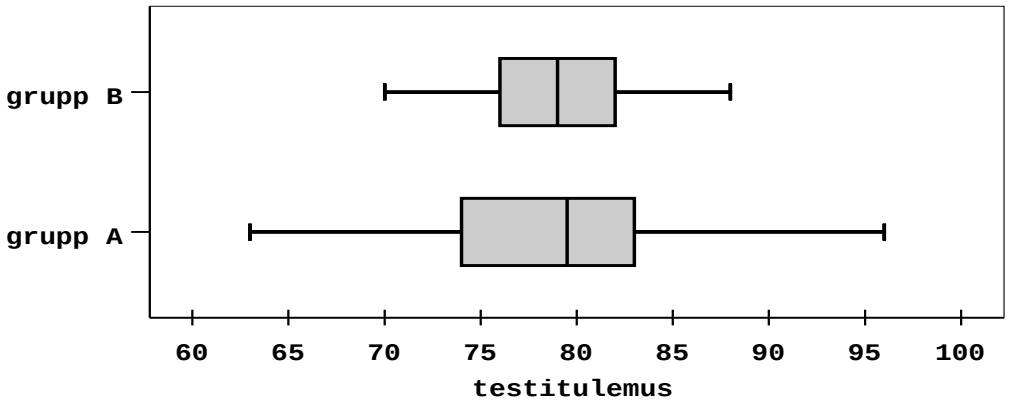
* * *

Kvartiilide vahe teises grupis on 2 palli. Ma arvan, et te ei kahtle, et kvartiilide vahe iseloomustab nende kahe grupi hinnete hajuvuse erinevust paremini kui jaotuse ulatus.

Lihtne on näha, et kvartiilide vahe määrab ära vahemiku, milles asuvad pooled valimi keskmisele lähedamal asuvad väärtused ning ulatuse ja kvartiilide vahe omavaheline võrdlemine annab meile pildi sellest, kuivõrd tugev on jaotuses keskele koondumise tendents. Kui kvartiilide vahe on ligikaudu võrdne poole ulatusega, siis tulemuste keskele koondumise tendents on olematu või väga väike ning tegemist on ühtlase jaotusega, kus igasuguseid väärtusi min ja max väärtuse vahel on enam-vähem võrdselt; kui aga kvartiilide vahe on oluliselt väiksem kui pool ulatust, siis on osad väärtused koondunud tihedalt ümber keskmise, kuid samas leidub üksikuid teistega võrreldes oluliselt madalamaid ja/või kõrgemaid väärtusi, mis jaotuse n.ö „välja venitavad“.

Jaotuse hajuvust saab iseloomustada ka graafiliselt histogrammi või KARPDIAGRAMMI[8] abil. Karpdiagrammil esitatakse üheaegselt mitu erinevat arvkarakteristikut: kvartiilid (seega ka mediaan) kujutatakse vertikaaljoontena, mille otspunktid ühendatakse horisontaaljoontega. Nii moodustub karp. Vurrude tippudeks valitakse valimi maksimaalne ja minimaalne väärtus ning ühendatakse need karbi serva keskpunktiga. Kui mõni väärtus asub mediaanist kaugemal kui poolteist kvartiilide vahet, siis märgib arvuti need diagrammile eraldi punktidena, et rõhutada nende erinevust teistest väärtustest. Eespool vaadeldud andmed kahe grupi testitulemuste kohta näeksid karpdiagrammil välja alljärgnevalt:

Kahe grupi testitulemuste jaotuste võrdlus (KARPDIAGRAMM)



Kõige sagedamini kasutatav hajuvuse näitaja on aga STANDARDHÄLVE. Nagu aritmeetiline keskmine, nii võtab ka standardhälve arvesse kõik vaatlustulemused. Kui meie vaatlustulemused on kõik ühesugused (nt kõik lapsevanemad hindasid mõtlemisoskuse arendamist kooli poolt väga tähtsaks st valisid 7-palli skaalal vastuseks 7), siis andmetes hajuvust ei ole ning mistahes hajuvuse näitaja peaks andma vastuseks 0. Paneme tähele, et juhul, kui andmetes hajuvus puudub, siis aritmeetiline keskmine on võrdne selle sama väärtusega, mida kõik vastajad valisid ehk ükski tulemus ei erine keskväärtusest. Tavaliselt on aga vaatlustulemused hajuvad ning üksikud tulemused erinevad (hälbivad) keskväärtusest enamal või vähemal määral. Standardhälve ongi selline arvkarakteristik, mis võimaldab meil öelda, kui palju üksikud tulemused grupi aritmeetilisest keskmisest (keskmiselt) erinevad. Mida suurem on hajuvus, seda suuremad on erinevused ning seda suurem on ka standardhälve.

Kumba jaotuse puhul allolevatest on teie arvates standardhälve suurem?

1) Kohuse- ja vastutustunne (viie lapsevanema vastused: 3 4 4 4 5) $\bar{x} = 4$

2) Aktiivsus, ettevõtlikus (viie lapsevanema vastused: 1 2 3 7 7) $\bar{x} = 4$

* * *

Väärtused teises reas on rohkem hajunud (st. nad erinevad ehk hälbivad keskväärtusest rohkem) kui esimeses reas. Seega võime arvata, et standardhälve on suurem teises reas olevate andmete puhul.

Vaatame nüüd, kuidas me seda arvude abil väljendada saaksime. Väärtused teises reas erinevad keskväärtusest alljärgnevalt:

Väärtus:	1	2	3	7	7
Erinevus $\bar{x}$ 'st:	- 3	-2	-1	+3	+3

[8] Vastavat diagrammi nimetatakse vahel ka pikemalt KARP-VURRUD-DIAGRAMMIKS.

Nüüd oleks meil vaja leida kui suur on keskmine erinevus keskväärtusest, kuid hälvete aritmeetilist keskmist me arvutada ei saa, sest negatiivsete ja positiivsete hälvete summa on alati = 0. Selleks, et pääseda mainitud tehnilisest raskusest tõstetakse kõik hälbed ruutu:

Hälve:	- 3	-2	-1	+3	+3
Hälve ruudus:	9	4	1	9	9

Saadud ruuthälvete aritmeetilist keskmist nimetatakse DISPERSIOONIKS:

$$Dispersioon = \frac{9+4+1+9+9}{5} = \frac{32}{5} = 6{,}4$$

Dispersioon on arvnäitaja, mida statistikas küllalt palju kasutatakse, kuid tal on andmetest ülevaate saamise kontekstis kasutamise jaoks üks tülikas puudus: kui vaatlustulemused (ja seega ka keskväärtus) olid näiteks ühikutes 'krooni', 'punkti' või 'millimeetrit', siis dispersiooni ühikuks oleks 'krooni ruudus' või 'millimeetrit ruudus'! Selliste ühikutega opereerimine ei oleks just kõige lihtsam ja mõistetavam. Selleks, et saada hälvet iseloomustavat suurust, mis oleks esialgsete andmetega samades ühikutes, leitakse ruutjuur dispersioonist - saadud näitajat nimetataksegi STANDARDHÄLBEKS:

$$Standardhälve = \sqrt{6{,}4} \approx 2{,}5 \text{ palli}$$

Mõneti tinglikult (ruutu tõstmise ja ruutjuure võtmise tõttu), aga sisuliselt siiski õigesti, võime saadud arvu tõlgendada nii, et keskmiselt erinesid viie lapsevanema arvamused grupi keskmisest arvamusest 2,5 palli võrra (tuletame meelde, et tegelikud erinevused olid -3, -2, -1, +3 ja +3 palli)[9].

Viime läbi samad arvutused jaotuse 1) jaoks:

Väärtus:	3	4	4	4	5
Erinevus $\bar{x}$ 'st:	- 1	0	0	0	+ 1
Hälve ruudus:	1	0	0	0	1

$$Dispersioon = \frac{1+0+0+0+1}{5} = \frac{2}{5} = 0{,}4$$

Standardhälve = $\sqrt{0{,}4} \approx 0{,}6$ palli

Nagu te arvata võisite, on esimese jaotuse standardhälve palju väiksem kui teise jaotuse puhul ning jääb alla ühe palli, sest üle ühe palli ei erinenud selle jaotuse puhul grupi keskmisest ju kellegi arvamus! Kui meil oleks tegemist suurema hulga andmetega (nt 68 lapsevanema arvamused), siis andmetele peale vaatamine (nagu antud „väikeses" näites) meile head ülevaadet vastuste hajuvusest ei annaks, kuid olles välja arvutanud, et vastuste standardhälve aktiivsuse ja ettevõtlikkuse tähtsuse hinnangute puhul on *s = 2,5* palli ning kohuse- ja vastutustunde tähtsuse hinnangute puhul ainult *s = 0,6* palli, saaksime kohe andmete kohta teha järelduse, et aktiivsuse ja ettevõtlikkuse arendamise tähtsuse osas läksid lastevanemate arvamused omavahel lahku e anti väga erinevaid hinnanguid, aga kohuse- ja

[9] Kuigi standardhälve esitatakse alati positiivse arvuna teame, et peale ruutjuure võtmist on vastus märgiga „$\pm$", mis tähendab antud kontekstis, et erinetakse nii alla kui üle keskmise.

vastutustunde tähtsust hindasid lapsevanemad väga sarnaselt. Pane tähele, et hajuvuse näitaja põhjal ei saa teha järeldust selle kohta, kumba hinnati tähtsamaks; selleks on vaja teada ka keskmist!

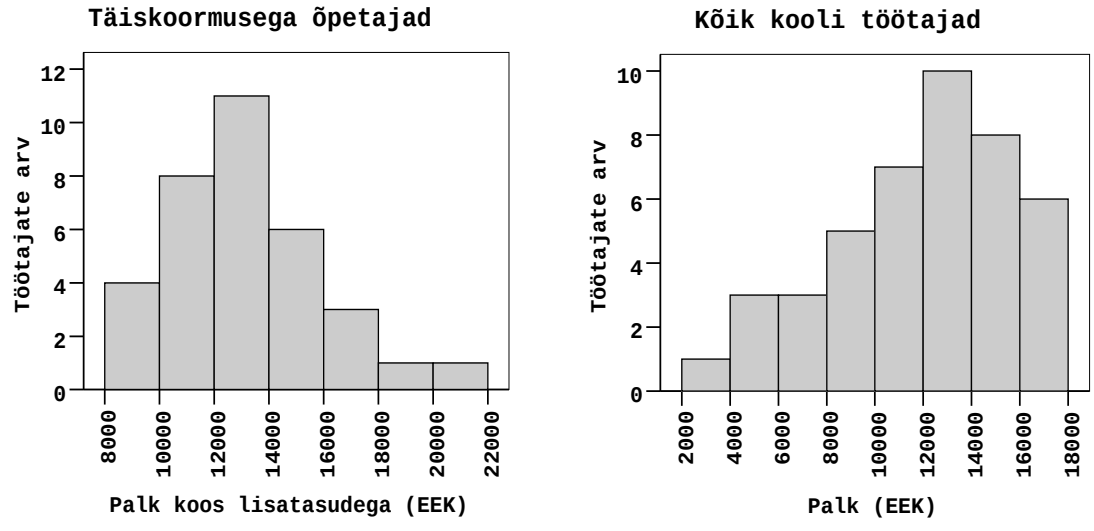
Aritmeetiline keskmine, standardhälve ja teised vaadeldud arvnäitajad võimaldavad meil ilmekalt kirjeldada arvtunnuste väärtuste jaotumist, kuid jaotuse kirjeldamisel on oluline osa ka joonistel ja graafikutel, mis võimaldavad luua parema ettekujutuse JAOTUSE üldisest KUJUST.

Ilmselt olete märganud, et paljude tunnuste puhul on tulemuste jaotus enam-vähem SÜMMEETRILINE s.t et kõige rohkem mõõtmistulemusi asub ulatuse keskosas ning liikudes ulatuse otspunktide poole mõõtmistulemuste hulk väheneb sarnaselt nii keskmisest madalamate kui keskmisest kõrgemate väärtuse puhul. Eelnevalt vaadeldud andmed testitulemuste kohta oli tüüpiline näide sellisest NORMAALJAOTUSELE lähedasest jaotusest. Selline sümmeetrilisus on statistiliste jaotuste puhul väga tavaline - eriti kui on tegemist bioloogiliste (või ka psühholoogiliste) nähtustega, kuid ta ei ole universaalne, ainuvõimalik – mida enam on nähtused, mida me uurime, mõjutatud ühiskonna ja kultuuriga seotud fenomenide poolt, seda tõenäolisem on, et andmete jaotus on ASÜMMEETRILINE.

Vaatleme kahte alltoodud näidet: vasakpoolsel joonisel on toodud andmed ühe väikese kooli täiskoormusega õpetajate palkade kohta. Näeme, et kõige tüüpilisemad palgad jäävad 12000 ja 14000 krooni vahele, kuid jaotus ei ole päris sümmeetriline, sest mõned teistega võrreldes kõrgemad palgad on jaotuse „välja venitanud“ paremale e matemaalilises mõttes positiivsele poole. Seetõttu räägitaksegi sellise jaotuse korral, et tegemist on POSITIIVSE ASÜMMEETRIAGA. Kui püüda mõelda „andmete taha“ ja küsida, miks selline asümmeetria tekkis, siis vastus antud näite korral peitub selles, et kahel õpetajal on teistega võrreldes rohkem täiendavaid ülesandeid – üks neist on direktori asetäitja õppetöö alal ning teine juhendab väga aktiivselt õpilaste huviringe. Samas parempoolsel joonisel on ühe teise väikese kooli kohta toodud samuti palkade jaotus, kuid nüüd on see „välja venitatud“ vasakule poole ja saame rääkida, et tegemist on NEGATIIVSE ASÜMMEETRIAGA. Millest võis siis antud juhul selline jaotuse kuju tekkida?

* * *

Tõepoolest, kuna joonise pealkirja järgi on selle kooli andmed esitatud mitte ainult täiskoormusega õpetajate vaid kõigi töötajate kohta, siis võib aimata, et jaotust vasakule venitavad tüüpilisest tunduvalt madalamad palgad on nt osakoormusega õpetajatel, koristajal, remonditöölisel, riidehoidjal jne.



Nägime, et jaotuse asümmeetriat saab tõlgendada s.t jaotuse kuju analüüsimine paneb meid oma andmete kohta sisulisi küsimusi esitama, mis aitab kaasa andmete olemuse paremale mõistmisele. Kuid jaotuse kujule on oluline tähelepanu pöörata ka seetõttu, et andmete põhjal arvutatud arvnäitajate tõlgendamine sõltub tihti jaotuse kujust. Näiteks nägime eespool, et sümmeetrilistes jaotustes on aritmeetiline keskmine ja mediaan suhteliselt sarnased ning seetõttu võib aritmeetilist keskmist tõlgendada kui valimit poolitavat näitajat, millest ligikaudu pooled tulemused/väärtused on madalamad ning ligikaudu pooled kõrgemad; samas asümmeetrilistes e ühele poole „välja venitatud" jaotustes on mediaan ja aritmeetiline keskmine erinevad ning seega ei jaga aritmeetiline keskmine valimit pooleks nagu ta seda sümmeetriliste jaotuste puhul teeb. Näiteks, on üsna oodatav, et suurtes läbilõikelistes valimites on sissetuleku või palkade jaotus positiivse asümmeetriaga ning aritmeetilisest keskmisest kõrgem sissetulek või palk on vaid ligikaudu ühel kolmandikul valimi liikmetest.

Asümmeetrilise jaotuse puhul võib erinevate keskmiste omavahelise paiknemise alati ette ennustada. Keskväärtus on moodist (e jaotuse tipust) [10] nihutatud jaotuse "saba" suunas ning mediaan asub nende kahe vahel. Mida suurem asümmeetria, seda suurem vahemaa jääb moodi, mediaani ja keskväärtuse vahele. Seda teades, saab teha ka vastupidiseid järeldusi: olles välja arvutanud aritmeetilise keskmise ja mediaani saame neid võrreldes pildi sellest, mis tüüpi jaotusega on tegu – kui mediaan ja keskväärtus on sarnased, siis on ilmselt tegemist sümmeetrilise jaotusega; kui keskväärtus on mediaanist märgatavalt madalam, siis viitab see jaotuse negatiivsele asümmeetriale ning kui keskväärtus on mediaanist märgatavalt kõrgem, siis positiivsele asümmeetriale. Jaotuse asümmeetria iseloomustamiseks on kasutusel ka vastav arvnäitaja, mida nimetatakse ASÜMMEETRIAKORDAJAKS.

Kokkuvõte

Käesolev peatükk algas tõdemusest, et uuringuid ei saa tihti läbi viia ilma meid huvitavate protsesside kohta andmeid kogumata. Andmete analüüsi tulemus saab aga usaldusväärne olla vaid juhul, kui kogutud andmete kvaliteet on kõrge. Seepärast tuleb juba enne andmete kogumist hoolikalt läbi mõelda, millistele küsimustele me andmete põhjal vastuseid tahame saada ning millisel viisil on kõige otstarbekam antud eesmärgist lähtuvalt andmeid koguda. Andmete kogumise instrumenti (nt küsimustikku) koostama asudes tuleb järgida lisaks sisulistele aspektidele ka tervet rida tehnilisemat laadi nõudeid ja põhimõtteid, mis aitavad tagada olukorra, kus vastaja motivatsioon sisuliselt õiget informatsiooni anda andmete kogumise käigus pigem tõuseb kui langeb ning kus nii vastaja kui andmete töötleja poolt kogemata tehtavate vigade võimalus on viidud miinimumini.

Mugava paindliku ja sügavuti mineva analüüsi tagamiseks on peale andmete kogumist mõistlik andmed sisestada arvutisse koostades lihtsa kuid põhireegleid järgiva struktuuriga algandmete tabeli. See esialgu ehk mõttetuna näiv lisatöö ja -aeg, mis kulub andmetabeli koostamiseks ja andmete sisestamiseks arvutisse, tasub end mitmekordselt ära andmete analüüsi etapis, kus andmete käsitsi kokku võtmine on väga ajamahukas isegi väikeste andmestike korral, kuid kus korraliku andmetabeli põhjal on arvuti abil mõne hetkega võimalik saada ülevaade oma andmetest mitme eri nurga alt ning leida vastused paljudele huvitavatele küsimustele.

Andmeid analüüsima asudes tuleb meeles pidada, et sugugi mitte kõik meetodid ei sobi samavõrra hästi kõigi meid huvitavate küsimuste või kõigi meie kasutuses olevate andmete korral. Kuigi statistilisi

[10] Arvtunnuste korral, millele on palju erinevaid väärtusi, on õigem siinkohal rääkida modaalsest intervallist e sellisest vahemikust, milles asub kõige enam väärtusi ja mitte üksikust moodiks osutunud väärtusest, sest niisuguste tunnuste puhul on mood praktikas ebastabiilne arvnäitaja.

meetodeid on väga palju, aitab esialgu õige meetodi valikul mõtlemine kahele suurele meetodite grupile, milleks on sagedustabelid ja arvnäitajad. Seetõttu võiks analüüsimeetodit valima hakates endalt küsida nt nii „Kas antud küsimuse ja andmete puhul on sobivam kasutada sagedusi või keskmisi ja teisi arvnäitajaid?“. Üldiselt võib meelde jätta, et sagedused „töötavad“ kõigi sellist tüüpi andmete puhul, kus tunnusel ei ole palju erinevaid väärtusi; arvnäitajatest enamus aga eeldab arvtunnuseid e andmeid, mis on mõõdetud/kogutud võrdsete vahedega skaalat kasutades. Võib juhtuda ka nii, et mõlema grupi meetodid on tehniliselt sobivad – sellisel juhul tuleb edasi mõelda sellele, kes on sihtrühm, kellele analüüsi tulemus esitatakse ning arusaadav peab olema? Pange tähele, et esmast analüüsi läbi viies olete „sihtrühmaks“ teie ise ja seega tuleb võimalike sobilike meetodite hulgast esmalt valida see meetod, mis teile endale kõige kiiremini ja mugavamini andmetest ülevaate ja vastused annab. Kui leiate midagi huvitavat, mis väärib laiemat tutvustamist, tuleb uuesti meetodid valikule mõtlema hakata ning võib juhtuda, et sõltuvalt sihtrühmast on nüüd vaja sama sisu esitamiseks valida mõni teine, sihtrühmale tuttavam / lihtsamini arusaadav / esitluskonteksti sobivam / jne meetod – seda muidugi ikka ainult nende meetodite hulgast, mis antud küsimuse ja andmete tüübi puhul tehniliselt korrektne kasutada on!

Lõpetuseks pidage meeles, et arvutit võib üsna julgelt usaldada arvutuste jms korrektsuse osas (eeldusel, et olete talle õiged käsud/juhised andnud), kuid tabelite ja diagrammide kujunduslik pool on see, mille osas ei maksa uskuda, et arvuti teile kohe parima võimaliku lahenduse pakub. Seega, tuleb selleks, et esitlus igati korrektne saaks, kasutajal endal pisut vaeva näha ja meeles pidada, et kõik, mis seondub kujundusega, peab teenima ainult üht eesmärki - äratama huvi ja usaldust esitatava info vastu ning aitama kaasa selle sisulisele mõistmisele!