ANDMETE KOGUMINE JA KORRASTAMINE

SISSEJUHATUS. ÜLDKOGUM JA VALIM

Arvandmete kogumine sai alguse koos esimeste riikide tekkega. Selleks, et ehitada püramiide ja kaevata niisutuskanaleid, pidi olema ülevaade tööjõust. Selleks, et määrata alamatele makse, pidi valitseja teadma, kui palju maad ja pudulojuseid on maksumaksjal. Andmeid rahvastiku kohta koguti juba 3. aastatuhandel e.Kr. Egiptuses ja Hiinas. Ka koraanis ja piiblis räägitakse rahvaloendustest. Umbes 1500 a. e.Kr. lasknud Mooses üle lugeda rahva meessoost liikmed. Kuningas Taavet korraldanud 210 aastat hiljem uue rahvaloenduse, et sõjaliste kokkupõrgete eel paremini hinnata võiduvõimalusi. Põhjalikke rahvaloendusi, kus lisaks pereliikmete arvule, soole, vanusele ja elukohale märgiti üles ka andmed varanduse kohta, korraldasid Rooma keisrid. Ka müüdis Jeesuse sünnist pidid Joosep ja Maarja minema Petlemma, et keiser Augustuse käsul lasta end üles kirjutada.

Statistika tänapäeva mõttes sai alguse 17.-18. sajandil ja ta tähendas algselt riigiteadust, milles kirjeldati rahvastikku, tööstust, armeed jmt. Tekkivad kindlustusfirmad vajasid täpset informatsiooni inimeste keskmise eluea, õnnetusjuhtumite sageduse aga ka majandusliku riski kohta.

Statistika on teadus, mis käsitleb arvandmete kogumist, töötlemist ja analüüsimist[1].

Arvandmeid võib koguda, töödelda ja analüüsida mitmeti. Kogumisel, töötlemisel ja analüüsimisel tehtavad vead andsid inglise looduseuurijale Francis Galtonile põhjuse teravmeelitsemiseks: "On olemas kolme liiki valesid: esiteks hädavale, mis on vabandatav, teiseks alatu vale, mida ei saa andestada, ja kolmandaks statistika.". Selleks et neid "kolmandat liiki valesid" oleks vähem, tuleb tunda *matemaatilist statistikat,* s.o. teadust, mis uurib seda, milliseid järeldusi antud andmetest saab teha ja milliseid järeldusi ei saa teha.

Matemaatiline statistika on matemaatika haru, mis uurib statistiliste andmete põhjal järelduste tegemise meetodeid.

Matemaatilise statistika üheks aluseks on tõenäosusteooria.

Statistikas on oluline uurimise objekt - *üldkogum*.

Üldkogum on kas looduse või ühiskonna nähtus või objektide hulk, mille kohta soovime teha teaduslikult põhjendatud järeldusi.

Üldkogumiks võib olla näiteks kogu Eesti elanikkond, kõik ühel aastal sündinud poisslapsed, kõik Tallinna tänavatel sõitvad autod, kõik Euroopa Liitu kuuluvad riigid, kõik lambivabrikus toodetud elektripirnid jmt.

Üldkogumil on palju teda iseloomustavaid tunnuseid, mis on ühel või teisel viisil mõõdetavad. Näiteks elektripirnil tema poolt kiirgatav valgushulk, tarbitav võimsus ja põlemisaeg; Eesti elanikul sünniaeg, sugu, haridustase, keskmine sissetulek, suhtumine mõnesse poliitilisse parteisse jne.

Üldkogumi uurimisel on kaks võimalust:

- uuritakse üldkogumi kõiki elemente
- uuritakse selle üldkogumi mingit osahulka ja tehakse selle osahulga uurimise põhjal järeldusi terve üldkogumi kohta.

Kuigi esmapilgul tundub, et uurida tuleks kogu üldkogumit, tehakse statistikas seda tegelikult harva. Paljudel juhtudel on üldkogumi uurimine seotud suurte kulutustega. Näiteks rahvaloenduse läbiviimine on kallis ja seetõttu korraldatakse neid keskeltläbi iga kümne aasta tagant, mõnes riigis sagedamini, mõnes vaesemas riigis harvem. Teiseks suhteliselt kalliks avaliku arvamuse uurimise võtteks on referendum. Referendumeid korraldatakse siis, kui on vaja saada kõigi elanike vastuseid riigi jaoks kõige olulisematele poliitilistele küsimustele.

Vahel on mõne toote uurimine seotud tema hävitamisega. Kui lambivabrikus kontrollitaks kõigi elektripirnide tööiga, siis ei jääks poodi saatmiseks ühtegi pirni järele. Seepärast kontrollitakse mitte kõiki pirne, vaid ainult juhuslikult valitud proovipartiisid.

Mõõtmiseks võetud üldkogumi osa nimetatakse valimiks.

Seega *statistika teeb järeldusi üldkogumi kohta valimi põhjal*.

Üks statistiliste andmete kogumise põhiprobleeme on järgmine:

Kuidas leida antud üldkogumist küllalt väike valim nii, et valimi kohta saadud tulemused kirjeldaksid ka üldkogumit piisavalt hästi?

Näide 1. Presidendivalimiste teises voorus jäid konkureerima kaks kandidaati: Armas Kuldsuu ja Ustav Tõelemb. Ajaleht "Eestimaa Suller" küsitles kahtkümmet Tallinna tänavatel liikuvat inimest, ja sai tulemuseks et

10 % vastanuist hääletab Kuldsuu poolt;
30 % Tõelembi poolt;
20 % läheb hääletama, aga ei tea veel, kelle poolt;
25 % ei lähe valima.

Ülejäänud küsitletud soovitasid küsitlejal "uttu tõmmata", "jalga lasta" või hiilisid mõnel muul viisil tüütavast pärimisest kõrvale.

Kuigi ajaleht "Eestimaa Suller" avaldas need tulemused ja ennustas, et Tõelemb kogub 75 % häältest ja Kuldsuu 25 %, ei võetud neid tulemusi tõsiselt. Miks?

Põhjus on väga lihtne - valim ei olnud küllalt arvukas.

Näide 2. Samade presidendikandidaatide populaarsust hindas Armas Kuldsuu parteikaaslaste poolt loodud avaliku arvamuse uurimisfirma "Amor". "Amor" korraldas telefoniküsitluse , milles helistati 3000-l järjestikusel telefoninumbril.
"Amor" sai tulemuseks et

52 % vastanuist hääletab Kuldsuu poolt;
13 % Tõelembi poolt;
30 % arvas, et ei lähe valima.

Ülejäänud vastanud arvasid, et lähevad valima, aga ei tea veel, kelle poolt hääletavad.

[1] Vahel nimetatakse statistikaks ka millegi kohta kogutud arvandmestikku.

Kuigi nüüd oli piisavalt palju küsitletuid, ei olnud valim moodustatud nii, et selle põhjal saaks järeldusi teha kogu Eesti kohta. Need 3000 järjestikust telefoninumbrit kuulusid ilmselt ühe keskjaama alla ja seetõttu ei näita need vastanud kogu Eesti hääletajate arvamust, vaid ainult ühe piirkonna telefonikasutajate oma. On võimalik, et ainult selles piirkonnas on Armas Kuldsuu populaarsem kui Ustav Tõelemb. Tulemused muutuksid usaldatavamaks, kui kõik telefoninumbrid oleks valitud täiesti juhuslikult. Aga ka siis on tulemuste usaldusväärsus liiga väike, sest kõik vastanud on telefoniomanikud, aga telefon kui tarbeese ei ole veel kättesaadav kogu ühiskonnale, vaid ainult jõukamale ühiskonna osale.

Seega

- *Valim peab olema küllalt arvukas.*
- *Igal üldkogumi objektil peab olema võrdne võimalus valimisse sattuda.*

Valimit võib moodustada juhuslikult, aga ka mingi kindla plaani alusel.

Juhusliku valimi saame, kui koostame üldkogumist mingi nimekirja ja võtame sealt juhuslikult välja uuritavad objektid. Täieliku juhuslikkuse saavutamiseks tuleks kasutada personaalarvuti või kalkulaatori juhuarvude generaatorit või spetsiaalseid juhuarvude tabeleid.

Näide 3. Lolliste linnavalitsus tahtis uurida linnakodanike suhtumist keskkütte hinna tõusu. Selleks valiti linna tänavate nimistust juhuslikult üks tänav ja küsitleti kõiki sellel tänaval elavaid inimesi. Tulemused olid ootamatud - küsitletud kiitsid keskkütte hinna tõusu heaks. Põhjus oli lihtne - sellel tänaval ei olnud üheski majas keskkütet. Järgmisel aastal otsustas linnavalitsus küsitleda elanikke suhtumist maa hinna tõusu. Et mitte korrata eelmine kord tehtud viga, otsustati seekord küsitleda iga perekonda, kes elab majas mille number lõpeb seitsmega, korteris number 16. Tulemused olid jälle ootamatud - elanikke jättis ka maa hinna tõus ükskõikseks. Põhjus oli lihtne: maa hind huvitab ennekõike eramajade omanikke, aga tüüpilises eramajas on alla 16 korteri. Kuigi valimisviis oli näiliselt juhuslik, ei saanud me kummalgi juhul juhuslikku valimit.

Juhusliku valimise kasutamisel on üks puudus. Kui me peame näiteks valima 1,4 miljoni Eesti elaniku hulgast täiesti juhuslikult välja 5000 isikut, siis tuleks meil ikkagi käia küsimusi esitamas peaaegu kõigis Eesti linnades, alevites ja valdades ja seetõttu kuluks uurimusele palju tööd, aega ja raha.

Planeeritud valimi korral saab uurimisele kuuluvat aega ja raha kokku hoida, aga tulemused võivad ikkagi tulla vajaliku täpsusega. Näiteks tüüpilise avaliku arvamuse küsitluse jaoks määratakse eelnevalt kindlaks, mitu linna ja mitu maaelanikku, mitu meest ja naist, mitu eestlast ja muulast küsitletakse. Need arvud määratakse vastavalt linna ja maaelanikkonna, meeste ja naiste, erinevate rahvuste protsentuaalsele jaotusele. Seejärel jaotatakse need arvud vastavalt linna või maakonna elanike arvule. Seejärel valitakse maakonnas juhuslikult välja linn, alev või küla, mille elanikke seekord küsitletakse. Igast linnast, alevikust või külast valitakse juhuslikult (nimekirja järgi) küsitletav. Kui küsitletavat ei ole kodus või kui ta keeldub kontakteerumast, siis on olemas reeglid, mille alusel leitakse teine küsiteldav.

Kui valim langeb kokku üldkogumiga, siis nimetatakse valimit *kõikseks valimiks*.

Ülaltoodud põhjustel kasutatakse kõikset valimit harva.

Kui statistilise uurimuse tegija uurib valimit, siis ta saab mõõtmise või küsitluse teel andmed, mis moodustavad *statistilise andmestiku*. Seda andmestikku võib hoida näiteks kartoteegina, milles iga uuritava objekti andmed on eraldi kaardikesel. Siiski on otstarbekam esitada nad tabelina, mille ridades on uuritavad objektid, veergudes aga nende objektide juures määratud tunnused. Sellist tabelit nimetatakse **objekt-tunnustabeliks**. Järgnev tabel on üheks näiteks objekt-tunnustabelist:

Riigikaitseakadeemiasse astujad pidid tegema läbi sisseastumiskatsed. Osa katsete protokollist nägi välja nii:

Perekonna ja eesnimi	Sugu	Vanus	Keskmine hinne	Testi punktid	Mundri nr.	Saapa nr.	Juhiload	Pikkus	Kaal
Tarmo Tank	m	20	3.5	81	56	44	A	187	95,5
Paul Püss	m	19	4.2	74	52	43	B	175	75,5
Miina Miin	n	23	3.5	52	52	40		170	8,05
Kai Kahur	n	19	4.9	62	46	38	B	164	65
Borja Bomm	m	18	3.0	100	60	47	C	199	120
Kusti Kuul	m	20	4.2	29	50	41		168	75
Tiia Tääk	n	26	4.0	48	44	36		18	65
Koit Kolt	m	19	3.8	75	52	42		183	75

Nagu näeme, võivad objekt-tunnustabelis kirjas olevad tunnused olla erinevat laadi. Tunnuseid, mille väärtusteks on arvud, nimetatakse **arvtunnusteks** ehk kvantitatiivseteks tunnusteks. Arvtunnused on näiteks pikkus, kaal, vanus, keskmine hinne, kinganumber, rahvaarv ja riigi pindala.

Tunnused, mille väärtuseks ei ole arvud, on **mittearvulised** ehk kvalitatiivsed tunnused. Mittearvulised tunnused on näiteks sugu, rahvus, haridus, juuste värv, perekonnaseis, ülaltoodud tabelis tunnus autojuhiload. Kvalitatiivse tunnuse väärtused ei ole arvulised, kuid neid võidakse märkida arvudega, et hõlbustada nende töötlemist arvuti abil.

Arvtunnused jagunevad omakorda

a) pidevateks;

b) diskreetseteks.

Pidev tunnus võib omandada kõiki reaalarvulisi väärtusi mingist piirkonnast.
Näiteks kaal, kasv, aeg ja temperatuur on pidevad tunnused.

Diskreetne tunnus võib omandada vaid üksteisest eraldatud väärtusi. Diskreetse tunnuse väärtused saadakse tavaliselt loendamise teel, näiteks perekonnaliikmete arv, õpilaste arv klassis.

Tunnuste jaotamine pidevateks ja diskreetseteks on mõnel juhul tinglik, sest sisuliselt pidev tunnus (vanus) võib osutuda diskreetseks väikese mõõtmistäpsuse tõttu (vanus täisaastates). Kui aga diskreetsel tunnusel on väga palju erinevaid väärtusi (näiteks erinevate perekondade aastasissetulekud kroonides), siis on mõnikord kasulik seda tunnust vaadelda kui pidevat. Andmete töötlemisel loetakse diskreetseks enamasti tunnus, millel on vähe väärtusi (mitte enam kui paarkümmend) ja need on täisarvud.

Mittearvulised tunnused jagunevad

a) nominaalseteks tunnusteks;

b) järjestustunnusteks.

Järjestustunnus on tunnus, mille väärtusi saab sisu põhjal järjestada. Näiteks küsimusele antud hinnangvastused (meeldib, olen ükskõikne, ei meeldi), hinded (väga hea, hea, keskpärane, puudulik).

Järjestustunnuse väärtusi võib esitada ka arvudena, aga nende arvude suhtearvudel ei ole mõtet. Näiteks hinne "4" pole kaks korda parem hindest "2", aga kaal "4 kg" on kaks korda suurem kaalust "2 kg".

Nominaaltunnused erinevad järjestustunnustest selle poolest, et neid ei ole väärtuse järgi mõtet järjestada. Näiteks rahvus, silmade värv, kutseala, parteilisus.

Nominaalsete tunnuste korral ainult loendatakse ühesuguseid väärtusi . Näiteks 1994. a. oli 10000 Eesti elanikust 6386 eestlast, 2897 venelast, 269 ukrainlast, 157 valgevenelast, 100 soomlast, 20 juuti, 24 tatarlast, 19 lätlast, 17 poolakat, 16 leedulast, 12 sakslast ja 83 muu rahvuse esindajat.

On olemas veel üks liik tunnuseid, mida mõned autorid loevad nominaal- aga mõned järjestustunnusteks. Need on n.n. *binaarsed* ehk *alternatiivsed* tunnused.

Binaarsel tunnusel on ainult kaks teineteist välistavat väärtust. Tüüpiline binaarne tunnus on sugu.

167. Sooviti uurida uue värvifilmi kvaliteeti. Kas on mõttekas kasutada kõikset valimit?

168. Õllevabrikant tahtis enne uue õllesordi "Haanja hapu" müügilepaiskamist uurida õllesõprade arvamust. Selleks korraldas ta õllebaaris "Roheline konn" tasuta degusteerimise. Kannutäie uut õlut sai igaüks, kes tellis kolm kannu alkoholivaba õlut "Võru vesine". Kirjelda valimit. Kas kõikne uurimus on mõeldav?

169. Mõtle välja veel mõni näide, kus terve üldkogumi uurimine on praktiliselt võimatu, küll aga saaks üldkogumit uurida valimi abil.

170. Kas Eesti keskkoolilõpetajate matemaatikateadmistest pildi saamiseks piisab ühe klassi õpilaste testimisest? Kas piisab ühe kooli abiturientide testimisest? Kuidas on määratud üldkogum? Kuidas koostada valim, nii et objektiivse pildi saaks vähimate kuludega?

171. Soovitakse uurida telesaadete populaarsust Eestis. Kuidas moodustaksid valimi?

172. Peeter väidab, et edetabeli "Tagumine paar välja" eesotsas olev laul on Eestis kõige populaarsem laul. Kas tal on õigus?
Kas "Kuku" raadio muusikasaate "Tagumine paar välja" muusikahindajate valimi moodustavad a) kõik Eesti kodanikud b) kõik Eesti noored c) kõik "Kuku" raadio kuulajad d) Kõik stuudiosse helistajad?

173. Mis on üldkogumiks ja valimiks
a) parlamendivalimistel; b) kohalikel valimistel?

174. Mis on üldkogum, valim ja mõõdetav tunnus
a) õhu temperatuuri määramisel ühes ilmajaamas, näiteks Kuusikul;
b) õhu temperatuuri määramisel kõigis Eesti ilmajaamades;
c) metsloomade ja lindude loendustel Eesti metsades?

175. Määra, mis tüüpi tunnused on järgmiste küsimuste vastused?
Nimi Sugu Sünniaasta Kodakondsus Haridustase
Töökoht Kuupalk Odaviske tulemus Kui tihti suitsetad?
Kuidas saab oma tööga hakkama Eesti Vabariigi peaminister?
Hinda kümne punkti skaalas ansambli "Vennaskond" meeldivust.

176. Sooviti uurida suitsetamise levikut noorte hulgas. Uurimust läbiviiv sotsioloog pakkus välja järgmised vastusevariandid:

a)	b)	c)
jah	ei	mitte ühtegi suitsu päevas
ei	harva	1-3 suitsu päevas
	peaaegu iga päev	4-10 suitsu päevas
	pidevalt	11-20 suitsu päevas
		üle ühe paki päevas

Sõnasta iga vastusevariandi puhul sobiv küsimus ning määra tunnuse tüüp.

177. Soovid uurida seda, millist muusikat Sinu koolikaaslased kuulaksid kooli raadiost vahetunni ajal. Mida küsiksid, et tehtav kahetunnine muusika-programm pakuks huvi võimalikult paljudele ja ei muutuks tüütavaks? Millist tüüpi vastuseid kasutaksid?

ANDMETE ETTEVALMISTAMINE TÖÖTLEMISEKS

Oletame, et küsitlus või mõõtmine on tehtud ja me oleme saanud andmetabeli. Nüüd võiks kohe hakata saadud tabelist järeldusi tegema, aga igaks juhuks uurime enne veel tabelit. Tabelis võib olla ka vigu ja täitmata lahtreid. Järgnevas tabeliosas märkame kahte tõenäolist viga. Tiia Tääk ei saa olla ainult 18 cm pikkune ja Miina Miin ei saa kaaluda 8,05 kg. Tiina Täägi pikkuse ülesmärkimisel võis mõni number jääda vahele - neiu pikkus võib olla 158 cm, 168 cm, 178 cm aga ka 180-189 cm.

Miina Miin kaalus tõenäoselt 80,5 kg (komaviga), aga võimalik on ka kahe vea koosesinemine (komaviga ja ärajäänud number).

Vigaseid väärtusi ei tohi asendada tõenäolise õige arvuga - niimoodi me võltsime andmeid ja teeme järeldusi võltsitud andmetest.

Vigaseid mõõtmistulemusi ei tohi ka asendada arvuga 0. Nüüd saaksime, et Miina kaalub 0 kg ja Tiia on 0 cm pikkune. See tulemus on samuti absurdne. Vigase mõõtmistulemusega lahtrisse tuleb jätta vastavasse kohta tühik või spetsiaalne puuduva väärtuse kood. Mõnes programmis on selleks arv -32767, mõnes -999.

Perekonna ja eesnimi	Sugu	Vanus	Keskmine hinne	Testi punktid	Mundri nr.	Saapa nr.	Juhiload	Pikkus	Kaal
Tarmo Tank	m	20	3.5	81	56	44	A	187	95,5
Paul Püss	m	19	4.2	74	52	43	B	175	75,5
Miina Miin	n	23	3.5	52	52	40		170	8,05
Kai Kahur	n	19	4.9	62	46	38	B	164	65
Borja Bomm	m	18	3.0	100	60	47	C	199	120
Kusti Kuul	m	20	4.2	29	50	41		168	75
Tiia Tääk	n	26	4.0	48	44	36		18	65
Koit Kolt	m	19	3.8	75	52	42		183	75

Puuduva väärtusega lahtreid võib edaspidises uurimuses kokku lugeda, aga neid ei või kasutada näiteks aritmeetilise keskmise arvutamisel. Asendades näiteks Tiia pikkuse -999-ga, liites kaheksa tabelis oleva noore pikkused ja jagades 8-ga, saaksime keskmiseks pikkuseks

$\overline{x} = (187+175+170+164+199+168-999+183):8 = 30{,}875$ cm, mis on absurdne.

Õige keskmise pikkuse leiame, kui jätame puuduva väärtuse välja:

$\overline{x} = (187+175+170+164+199+168+183):7 = 178$ cm.

Vahel on tark mõõtmis- ja küsitlustulemused enne edasist uurimist *kodeerida*. Vastuvõtukomisjon tahab vastuvõtutesti tulemuse ja keskkooli keskmise hinde summa alusel määrata õppurikandidaatide paremusjärjestuse. Kui liita testi punktid keskmisele hindele, siis on saadud summas keskmise hinde osatähtsus võrreldes testi tulemustega tühine. Seepärast tuleb kasutada mingit kodeerimiseeskirja, mis teisendaks testi punktid hindeks. See võiks olla näiteks järgmine:

Punktide arv	0 - 20	21 - 40	41-60	61-80	81-100
Hinne	1	2	3	4	5

Selline punktide kodeerimine hindeks ei ole ainuvõimalik, sest tulemus sõltub hindamisskaala valikust (näiteks Saksamaal on vastupidine süsteem - kõige parem hinne on 1 ja halvim 5, Prantsusmaal hinnatakse aga 20 palli süsteemis).

Kodeerimine on tunnuste väärtuste hulga teisendamine, milles igale tunnuse esialgsele väärtusele seatakse vastavusse üks uus väärtus - kood.

Järjestustunnused tuleb töötluseks üldjuhul kodeerida. Koodidena kasutatakse tavaliselt arve. Järjestustunnuse kodeerimisel on mõistlik säilitada sisuline järjestus. Näiteks vastusevariandid hinnangküsimusele saab kodeerida nii:

meeldib väga	meeldib	olen ükskõikne	ei meeldi	üldse ei meeldi
5	4	3	2	1

Võimalikud on ka muud kodeerimisviisid. Võib kasutada vastupidist skaalat (meeldib väga - 1, ... ,üldse ei meeldi - 5), skaala võib alata 0-st jne.

Kodeerimiseeskirjad võivad kaotada osa teavet. Testis kogutud punktisumma andis meile rohkem informatsiooni, kui testi tulemus hindena 5 pallises skaalas.

Aga alati ei olegi meil järelduste tegemiseks nii palju informatsiooni vaja kui testi tegemisel küsisime. Olgu meil näiteks tunnuse *elukoht* erinevad kodeerimiseeskirjad.

Elukoht	Suurlinn	Linn	Väikelinn	Alev	Alevik	Maa
A	1	1	1	1	2	2
B	6	5	4	3	2	1

Kodeerimiseeskiri A on küll korrektne, aga kaotab osa teavet, nimelt erinevuse suur- ja väikelinna vahel. Mõne küsimuse lahendamisel (näiteks maa- ja linnaelanike protsendi määramisel) pole aga seda teavet vaja.

Nominaaltunnuseid võib kodeerida arvudena, aga ei tohi töödelda arvudena.

Kui näiteks ülaltoodud sisseastujate tabelis juhilubade olemasolu kodeerida nii:

Lube pole	A	B	C	D
0	1	2	3	4

Kui nüüd leida tulemuste aritmeetiline keskmine, siis saame

$\overline{x} = (1+2+0+2+3+0+0+0):8 = 1$. Seega keskmisel sisseastujal peaksid olema olemas mootorratturi load, aga tegelikult on need ainult Tarmol.

Andmetöötluse aluseks on andmetabel ehk objekt-tunnustabel, milles osa andmeid võivad olla kodeeritud kujul. Et see tabel oleks üheselt mõistetav töötlejale ja ka tabeli koostajale, selleks peab tabelile lisama *andmekirjelduse*.

Andmekirjelduses on

1) tunnuste nimed ja nimede tähendused

2) tunnuste tüübid;

3) kodeerimiseeskirjad;

4) arvtunnuste korral ka mõõtühikud.

Tunnuse nimi on tunnuse nimetus. Näiteks sugu, vanus, elukoht. Vahel võib tunnuse nimetus olla pikem tekstilõik, näiteks *ühes kuus tehtavad kulutused toidule, alkohoolsetele jookidele ja tubakatoodetele.* Kui nii pikka tekstilõiku kasutada tabeli päises veeru kohal, siis läheb veerg, milles on tavaliselt kolme kuni viiekohaline arv liiga laiaks ja selliseid veerge sisaldav tabel ei mahu laiust pidi paberile, arvutiekraanile ja printerile. Järelikult on mõistlik tunnuse nimena kasutada mingit lühemat tekstilõiku, näiteks *toidukulu*. Tunnust, mille nimeks on *kulutused tööstuskaupadele, teenustele, säästudeks ja maksudeks* võiks nüüd nimetada näiteks *muudkulud*. Sellised lühinimetused lausa nõuavad, et neid oleks pikemalt seletatud andmekirjelduses. Muidu võib näiteks kümne aasta pärast seesama tabeli koostaja arvata et *toidukulu* tähendab päevast energeetilist toiduvajadust kilokalorites, *muudkulud* aga päevast alkoholi ja tubakavajadust.

Et objekt-tunnustabelisse mõõtühikuid ei märgita, siis on loomulik ka see, et andmekirjelduses on arvtunnustel kirjas ka mõõtühikud.

Kaasajal töödeldakse statistiliselt andmeid enamasti arvuti abil ja kasutatakse vastavaid tarkvarapakette. Arvutiprogrammi võimalused määravad ära kasutatavate tunnuste tüübid. Need ei lange päriselt kokku meie poolt varem tehtud tunnuste jaotusega (arvtunnused ja mittearvulised tunnused, pidevad tunnused ja diskreetsed tunnused, binaartunnused, nominaal- ja järjestustunnused).

Arvutiprogrammis võib tunnus olla

a) tekstilõik; b) naturaalarv; c) fikseeritud pikkusega reaalarv;

d) kuupäev; e) tõeväärtus (1 või 0).

Viimast tüüpi kasutatakse binaartunnuste (näiteks sugu) märkimiseks.

On loomulik, et andmekirjeldus sisaldab ka tunnuste kodeerimisreegleid.

Selline andmekirjeldus aitab neid andmeid kasutada ka mõne aasta pärast, kui näiteks tahetakse võrrelda erinevate aastakäikude tulemusi, aga keegi ei mäleta enam, mida tähendasid lühendatud tunnuste nimed ja kuidas lahti mõtestada arusaamatuna tunduvat kodeeringut.

Teema lõpetuseks kõige olulisem. Selleks et koostada andmestikku, peab olema selge, *mida me tahame uurida*. *Me peame teadma, millistele küsimustele tahame andmestiku abil vastused leida*. Andmeid koguda ainult selleks, et andmeid koguda, ei ole mõtet.

Andmete kogumine ja korrastamine on vajalik eeltöö andmete statistilisele töötlemisele. See eeltöö koosneb järgmistest etappidest:

1) *probleemi püstitamine ja üldkogumi määratlemine*;
2) *mõõdetavate tunnuste ja mõõtmistäpsuse määramine*;
3) *valimi suuruse määramine ja valimi moodustamine*;
4) *tunnuste väärtuste mõõtmine valimil*;
5) *kodeerimiseeskirjade fikseerimine*;
6) *andmekirjelduse lisamine andmestikule*.

178. Veidrikust füüsikaõpetaja mõõdab õpilaste teadmisi 10 palli skaalas. Mõtle välja sõnalised hinnangud tema hinnetele.

179. Poliitikute populaarsuse ja tuntuse küsimusel on viis erinevat vastust:
usaldan täielikult, usaldan, ei usalda, üldse ei usalda, ei tunne sellist.
Koosta kolm erinevat kodeerimiseeskirja.

180. Mõtle välja kahed võimalikud vastusevariandid küsimusele "Kuidas suhtud noormeeste kohustuslikku sõjaväeteenistusse?". Mõlemal juhul pane kirja korrektne kodeerimiseeskiri ja määra tunnuse tüüp.

181. Õpetajad väidavad, et õpilastel on vaba aega laialt, aga nad kasutavad seda ainult logelemiseks. Teie väidate vastupidist, s.t. et olete nii koormatud koduste ülesannetega, et vaba aega ei jää. Koosta küsimustik, mis aitaks välja selgitada, milleks kulub õpilase aeg keskmisel koolipäeval. Kui palju on tal vaba aega?[1]

182. Seni kooli toitlustamist korraldanud firma läks pankrotti. Osa õpilasi tahaks koolis suppi või praadi süüa, osa lepiks puhvetist saadava kohvi ja saiakestega, osa näriks vahetunni ajal kodunt kaasa toodud võileibu. Sina kui aktiivne abiturient pead korraldama küsitluse, millest selguks: "Millist toitlustamisviisi õpilased eelistaksid ja kui palju on nende vanemad valmis kulutama raha lapse toitlustamisele koolis. Koosta küsimustik ja mõtle välja küsimused, vastusevariandid küsimustele, kasutatavad tunnused, nende tüübid ja kodeerimisreeglid.

183. Koguge ühiselt oma klassi kohta andmestik, mis sisaldaks kindlasti järgmisi tunnuseid: eesnimi, sünniaasta, -kuu, -päev, vanus, kasv, kaal, jalanumber, lemmikloom, lemmikansambel, matemaatika hinne, füüsika hinne, eesti keele hinne. Andmestik on hädavajalik edaspidiseks tööks.

[1] See ülesannne, aga ka järgmised kaks ülesannet ning ka ülesanne 177. on mõeldud iseseisva või grupiviisilise projektülesandena, mis tuleks lahendada nädala-kahe jooksul. Selliseid ülesandeid vaatleme edaspidigi, nende läbitegemine annab vajaliku kogemuse statistika paremaks mõistmiseks.

ANDMETÖÖTLUS

VARIATSIOONRIDA. SAGEDUSTABEL

Näide 1. Kaitseministeerium tahab kotakombinaadilt tellida kõigi sõdurpoiste jaoks saapad. Ministeerium teab küll ligikaudselt sõdurite üldarvu, aga ta ei tea, kui palju on vaja tellida erineva suurusega saapaid. Võiks muidugi uurida üldkogumit, aga sel pole erilist mõtet, sest järgmisel aastal on sõjaväes juba uued poisid. Lihtsam on moodustada paari sõjaväeosa juhuslikult võetud sõduritest valim ja selle abil selgitada saabaste suhteline vajadus vastavalt numbrile. Ühe rühma sõdurpoisid kandsid järgmise suurusega saapaid:

43 41 42 43 44 44 40 43 42 43 44 42 43 46 44 40
45 42 43 41 42 43 44 43 41 42 41 43 42 44 41 42
43 45 44 46 40 41 43 44

Selline arvude järjestus on *statistiline rida*. Statistilise rea *maht* on elementide arv selles reas. Et siin on toodud 40 sõduri saapanumbrid, siis rea maht on 40.

Et veltveebel Mesipuu ei viitsinud hakata neis andmetest erineva suurusega saabaste kandjaid kokku lugema, siis käskis ta poistel võtta ritta mitte pikkuse, vaid hoopis saapa numbri järgi. Tulemuseks oli järgmine andmete rida

46 46 45 45 44 44 44 44 44 44 44 44 43 43 43 43
43 43 43 43 43 43 43 42 42 42 42 42 42 42 42 41
41 41 41 41 41 40 40 40

Tulemuseks sai veltveebel Mesipuu tunnuse (saapanumber) järjestatud väärtuste rea ehk *variatsioonrea.*

Kasvavalt või kahanevalt järjestatud tunnuse väärtuste rida nimetatakse variatsioonreaks.

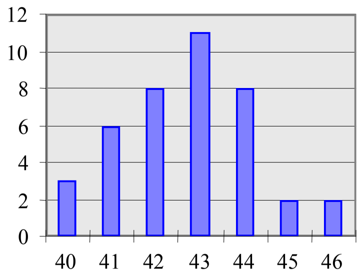
Saadud variatsioonrea põhjal koostas veltveebel Mesipuu järgmise tabeli.

Saapa number	40	41	42	43	44	45	46
Kandjaid	3	6	8	11	8	2	2

Sellist tabelit nimetatakse *sagedustabeliks*.

Sagedustabel näitab, mitmel korral antud tunnus saab antud väärtuse.

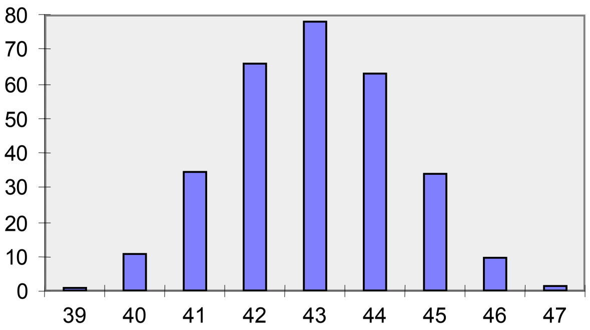
Sagedustabelist graafilise ülevaate saamiseks koostatakse tulpdiagramm.



Selle tabeli põhjal oli veltveebel Mesipuule selge, et ilmselt tuleb keskmise suurusega saapaid (numbrid 42-44) tellida rohkem kui väikesi ja suuri. Et mõõtmisi on tehtud vähe, siis seda, kui palju tuleb tellida erinevas suuruses saapaid, veel otsustada ei saa. Kaitseministeeriumi referent Kaelapuu laskis aga teha analoogilised uuringud ka mõnes teises väeosas, võttis kokku saadud tulemused ja sai järgmise sagedustabeli

Saapa number	39	40	41	42	43	44	45	46	47
Sagedus	1	11	35	66	78	63	34	10	2

Tabelis on 300 noore sõjamehe saapanumbrid. Joonistades välja sellele tabelile vastava tulpdiagrammi, sai ta järgmise pildi:



Tulpdiagrammilt ei ole näha, mitu protsenti peab tellima mingi suurusega saapaid. Referent Kaelapuu otsustas viimasele tabelile liita veel ühe veeru, kus on kirjas, mitu protsenti saapakandjaist kannab antud numbriga saapaid:

Saapa number	39	40	41	42	43	44	45	46	47
Sagedus	1	11	35	66	78	63	34	10	2
Sagedus %-des	0.33	3.67	11.67	22	26	21	11.33	3.33	0.67

Saadud tabelit nimetatakse *sagedus-jaotustabeliks*.

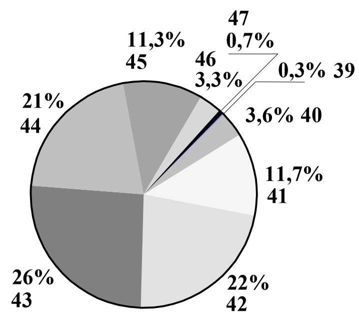
Kui jätta viimasest tabelist ära keskmine rida, siis saame *jaotustabeli*.

Jaotustabel näitab tunnuse väärtuste suhtelist esinemissagedust.

Suhteline esinemissagedus saadakse esinemissageduse jagamisel kõigi mõõdetud objektide arvuga. Vahel esitatakse suhteline esinemissagedus protsentides.
Tunnuse *saapa number* jaotustabel tema tüüpilisel kujul näeb välja nii:

Saapa number	39	40	41	42	43	44	45	46	47
Jaotus	0.0033	0,0367	0,1167	0,22	0,26	0,21	0,113	0,033	0,0067

Jaotustabelit saab esitada ka sektordiagrammi abil. Sektordiagramm annab hea ülevaate terviku jagunemisest, aga ta ei sobi absoluutarvude esitamiseks. Arv- ja järjestustunnuste puhul on eelistatavam tulpdiagramm, sest selle horisontaalteljel on näha tunnuse väärtuste järjestus.



Nominaaltunnuste puhul ei ole järjestus tähtis ja seetõttu kasutatakse rohkem sektordiagrammi.

Sagedustabelit saab kasutada ka pideva tunnuse erinevate väärtuste paiknemise ja esinemise sageduse uurimiseks.

Näide 2. Veltveebel Mesipuu, jefreitor Küünarpuu ja paljud teised alamväelased mõõtsid väeosades kutsealuste pikkused ja saatsid tulemused Kaitseministeeriumi referent Kaelapuule. Kaelapuul oli andmeid vaja selleks, et määrata, kuidas jagunevad sõdurpoisid pikkuste järgi. Tal oli nimelt vaja teada, kui palju peab erinevat kasvu mundreid tulevaks aastaks tellima. Kutsealuse pikkust kui tunnust võib lugeda pidevaks. Mundri kasv kui tunnus on aga diskreetne - sest rätsepatöökojad õmblevad masstoodangut ainult teatud kasvude kaupa. Õmblusvabrik soovitas Kaelapuul jagada kõik pikkused kaheksaks vahemikuks ja loendada igasse vahemikku sattunud pikkused. Kaelapuu tegigi seda nõutud viisil ja sai tulemuseks sagedustabeli:

Pikkus	-157	158-164	165-171	172-178	179-185	186-192	193-199	200-
Sagedus	0	17	50	83	84	48	16	2

Esita see sagedustabel tulpdiagrammina ja koosta vastav jaotustabel.

Sagedustabeli moodustamiseks jaotatakse pideva tunnuse kõikvõimalike väärtuste hulk ühisosata vahemikeks ehk **klassideks.** Vahemiku otspunkte nimetatakse *klassipiirideks*. Klassipiirideks valitakse enamasti täisarvud, kusjuures otsmised klassid võivad olla ka lahtised, s.t. vähima klassi alumist ja suurima klassi ülemist piiri määratud ei ole.

184. Moodusta leheküljel 54 antud tabeli andmete põhjal tunnuste *vanus* ja *keskmine hinne* variatsioonread.

185. Moodusta oma klassi andmestiku põhjal tunnuste *pikkus*, *kaal* ja *jalanumber* variatsioonread.

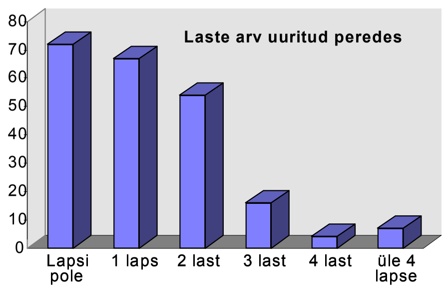
186. Moodusta oma klassi andmestiku põhjal tunnuste sagedus ja jaotustabelid

a) *sünnikuu* b) *vanus* c) *jalanumber* d) *eesti keele hinne.*

e) *pikkus* f) *lemmikloom* g) *poiste kaal*

187. Joonesta oma klassi andmestiku põhjal eelmises ülesandes toodud tunnuste tulpdiagrammid.

188. Tulpdiagramme võib kujutada ka kolmemõõtmeliselt. Kas kõrvalolev tunnus on pidev või diskreetne tunnus?



KESKVÄÄRTUS. MEDIAAN. MOOD

Alati pole andmestiku iseloomustamiseks vaja esitada kõiki statistilisi andmeid vaid piisab üksikutest näitajatest - *karakteristikutest.* Tuntuimad karakteristikud on *keskmised* ja *hajuvusmõõdud*. Käesolevas teemas tutvume lähemalt keskmiste leidmisega.

Keskmised, nagu nimetuski ütleb, väljendavad antud tunnuse mingit *keskmist väärtust*, mille ümber tunnuse väärtused paiknevad.

Tuntuimad keskmised on *geomeetriline keskmine, harmooniline keskmine, ruutkeskmine, keskväärtus, mediaan ja mood.* Meie käsitleme siinkohal põhjalikumalt keskväärtust, mediaani ja moodi.

Tunnuse keskväärtuseks on tunnuse väärtuste aritmeetiline keskmine.

Keskväärtust tähistatakse $\bar{x}$.

Olgu x_1 vaadeldava tunnuse väärtus esimese objekti korral, x_2 teise objekti korral jne ning n olgu mõõdetud objektide arv.

Aritmeetiliseks keskmiseks nimetatakse tunnuse kõigi väärtuste summa ja objektide arvu jagatist.

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}}$$

Seda tulemust saab Leonhard Euleri poolt kasutusele võetud *summamärgi* Σ (kreeka tähestiku suur täht sigma) abil[1] esitada lühemalt:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Summamärgi abil tähistatakse lühemalt ühelaadsete suuruste summat.

Näiteks $a_k + a_{k+1} + \ldots + a_n = \sum_{i=k}^{n} a_i$, kus a_i on summa üldliige, i summeerimisindeks, ning k ja n vastavalt summeerimise alumine ja ülemine raja.

Näiteks summa $1+2+3+\ldots+50 = \sum_{i=1}^{50} i$, aga summa $\frac{1}{3} + \frac{1}{9} + \frac{1}{27} + \ldots = \sum_{i=1}^{\infty} \frac{1}{3^i}$.

Seega keskväärtus $$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .}$$

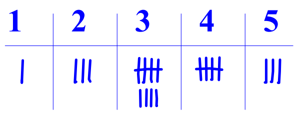
[1] $x_1 + x_2 + x_3 + \ldots + x_n = \sum_{i=1}^{n} x_i$.

Vaatame veelkord pealkirja “Andmete ettevalmistamine töötlemiseks” järel toodud tabelit. Tunnuse *pikkus* väärtused olid: 187, 175, 170, 164, 199, 168, 18 (mõõtmis-viga), 183.
Keskväärtuse leidmiseks jätame vigase lähteandme vaatluse alt välja, liidame ülejäänud arvud ja jagame tulemuse 7-ga:

$$\bar{x} = \frac{187+175+170+164+199+168+183}{7} = 178 \text{ cm}.$$

Objektide arv, millega jagame, on võrdne valimi mahuga ainult siis, kui ühelgi objektil vaadeldava tunnuse väärtus ei puudu. Kui andmestikus on tühikuid, siis on objektide arv nii mitme objekti võrra väiksem, kui mitu mõõtmistulemust vaadeldaval tunnusel puudub.

Näide 1. Keemiaõpetaja Mensuur tegi ühel päeval kolmes paralleelklassis korraga kontrolltöö. Ta parandas tööd ära ja tahtis teada, milline on kõigi tööde keskmine hinne. Ta hakkas tulemusi kalkulaatoril liitma, aga et liita oli peaaegu sada arvu, siis juhtus tal äpardusi: kord jukerdas klaviatuur, nii et hinde 2 asemel sisestas ta 22; siis vajutas ta kogemata liitmise klahvi asemel jagamise klahvile; siis tuli see tüütu geograaf Gloobus oma tobedate anekdootidega ja ta kaotas oma järje. Iga kord tuli otsast alata. "Kas kuidagi lihtsamalt ei saa?", küsis ta matemaatikaõpetajalt. "Saab küll, kasuta sagedustabelit" vastas matemaatik Abstsiss ja näitas Mensuurile, kuidas koostada järgmist tabelit:



Niiviisi jätkanud ja lugenud kokku lahtrisse sattunud kriipsukesed, sai proua Mensuur sagedustabeli:

Hinne	1	2	3	4	5
Sagedus	3	24	39	23	6

"Mis ma nüüd edasi teen?", küsis keemiaõpetaja. "Nüüd korrutad iga hinde tema esinemissagedusega ja liidad saadud tulemused. Saadud summa jagad hinnete arvuga", vastas Abstsiss.

Teinud nii, sai keemik järgmise tulemuse:

$$\bar{x} = \frac{1\cdot 3+2\cdot 24+3\cdot 39+4\cdot 23+5\cdot 6}{95} = \frac{290}{95} \approx 3{,}05.$$

Kui objekte on palju, siis on mõistlik koostada sagedustabel ja keskväärus leida järgmise valemi abil:

$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot f_i\ .}$$

Siin n on statistilise rea objektide arv, f_i on väärtuse x_i esinemise absoluutne sagedus (kordade arv).

Viimast valemit on mõistlik kasutada ka siis, kui tahame leida mõne pideva tunnuse keskväärtust.

Näide 2. Referent Kaelapuul oli sagedustabel sõdurpoiste pikkustest:

Pikkus	-157	158-164	165-171	172-178	179-185	186-192	193-199	200-
Sagedus	0	17	50	83	84	48	16	2

Selle tabeli põhjal saab määrata antud valimi pikkuse keskväärtuse. Et erinevad tunnuse väärtused kujutavad endast vahemikke, siis kogu vahemikku iseloo-mustavaks väärtuseks võtame vahemiku keskpunkti. Näiteks vahemiku 158-164 keskpunktiks on 161. Lõpmatute lahtiste vahemike (esimene ja viimane) asemel vaatleme vahemikke 151- 157 ja 200-206. Nende vahemike keskpunktid oleksid 154 ja 203. Seega saame

$$\bar{x} = \frac{154\cdot 0+161\cdot 17+168\cdot 50+175\cdot 83+182\cdot 84+189\cdot 48+196\cdot 16+203\cdot 2}{300} = \frac{53564}{300} \approx 178{,}5.$$

Eespool veendusime, et nominaaltunnuseid ei ole mõtet töödelda arvudena, seega ka nominaaltunnuse keskväärtust pole mõtet leida.

Keskväärtusel kui keskmisel on ka teatud puudused. Arvutamise tulemusena saadud väärtus ei pruugi ise olla üks tunnuse väärtustest. Näiteks keskmine hinne ei ole tavaliselt hinne, s.t. naturaalarv, vaid on ratsionaalarv.

Keskväärtus kui karakteristik iseloomustab üldkogumit halvasti siis, kui tunnuse väärtuste hulgas on üks või mitu väärtust, mis on ülejäänud väärtustest palju suuremad või palju väiksemad.

Näide 3. VI^B klassis on kümme poissi, neist 9 on tavalised kuuenda klassi poisid, kümnes on aga kolm aastat istuma jäänud Leopold Lohemadu, kes on juba täismehe kaalu ja kasvuga. Poiste pikkused (cm) kasvamise järjekorras on

148, 149, 150, 151, 152, 152, 153, 153, 154, 188.

Kui nüüd leida nende poiste kasvu keskväärtus, siis saame tulemuseks

$$\bar{x} = \frac{148+149+150+151+152+152+153+153+154+188}{10} = 155.$$

Seega on nende andmete korral üheksa poisi pikkused keskväärtusest väiksemad ja ainult ühe pikkus on suurem. VI^B klassi poiste keskmine pikkus on tõesti 155 cm, aga keskmine selle klassi poiss on siiski sellelst pikkusest lühem. Kogumi paremaks iseloomustamiseks tuleb appi võtta mediaan.

Mediaan on arv, millest suuremaid ja väiksemaid väärtusi on variatsioonreas ühepalju.

Mediaani tähistatakse statistikas sümbolitega *Me*.

Kui variatsioonreas on paaritu arv liikmeid, siis mediaaniks on selle rea keskmine liige. Kui variatsioonreas on paarisarv liikmeid, siis mediaaniks on kahe keskmise liikme poolsumma.

Näiteks VI^B klassi poiste pikkuse variatsioonrea

148, 149, 150, 151, $\boxed{152}$, $\boxed{152}$, 153, 153, 154, 188

mediaaniks on kahe keskmise liikme poolsumma, s.t. $Me = \frac{152+152}{2} = 152$ cm.

Leiame õpetaja Mensuuri poolt korraldatud keemia kontrolltöö hinnete mediaani.

Hinne	1	2	3	4	5
Sagedus	3	24	39	23	6

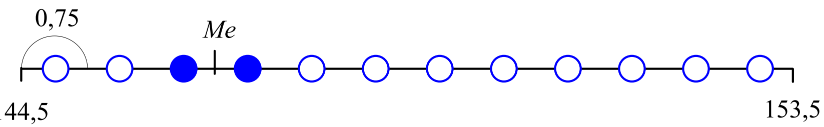
Et objekte on sagedustabelis 95, siis mediaaniks on sellele tabelile vastava variatsioonrea keskmine väärtus, s.o. 48. väärtus. Suuruselt 48. väärtus kuulub tabeli järgi arvutades (3+24+21) kolmandasse tulpa. Seega mediaan on 3.

Tülikam on mediaani leidmine siis, kui sagedustabelis on pideva tunnuse väärtused.

Leiame tunnuse *pikkus* mediaani järgmise sagedustabeli põhjal:

Pikkus	118-126	127-135	135-144	145-153	154-162	163-171	172-180
Sagedus	3	5	9	12	5	4	2

Et tabelis on 3+5+9+12+5+4+2 = 40 objekti, siis mediaan on 20. ja 21. objekti aritmeetiline keskmine. Leiame need objektid. Variatsioonrea 20. ja 21. objekt kuuluvad vahemikku 145-153 ja on selles vahemikus vastavalt 3. ja 4. objekt. Selle vahemiku täpsed piirid on vastavalt ümardamisreeglitele 144,5 ja 153,5 cm. Oletades, et mõõtmistulemused paiknevad vahemikus *ühtlaselt* (s.t. nende vahed on ühesugused), ja kasutades seda, et vahemiku laius on 9, saame et mõõtmistulemuste vahed selles vahemikus on $\frac{9}{12} = 0{,}75$ ühikut. Kujutades seda vahemikku joonisel, näeme, et mediaan asub vahemiku täpsest alumisest piirist $3 \cdot 0{,}75 = 2{,}25$ cm kaugusel.



Seega tuleb mediaani leidmiseks vahemiku täpsele alumisele piirile lisada 2,25.

Seega $Me = 144{,}5 + 2{,}25 = 146{,}75$.

Mediaani arvutamisel viimases näites eeldasime, et mõõtmistulemused paiknevad vahemikus ühtlaselt. Paraku me ei või seda kindlasti väita. Seetõttu saime me tulemuseks mitte mediaani täpse väärtuse, vaid hoopis *mingi ligikaudse hinnangu* mediaani väärtuse jaoks. Hinnangutega oli meil tegu ka sõdurite pikkuse keskväärtuse määramisel sagedustabeli abil.

Veel üldisemalt rääkides, kui me leiame keskväärtuse, mediaani või mõne muu karakteristiku valimi andmete põhjal,ja meil ei ole tegu kõikse valimiga, siis me ei saa väita, et et leidsime üldkogumi keskväärtuse, mediaani või mõne muu karakteristiku. Tegemist on ainult üldkogumi selle karakteristiku hinnanguga.

Nii nagu keskväärtust, nii pole ka mediaani mõtet leida nominaaltunnuste korral.

Erinevalt keskväärtusest ei ole mediaan seotud tunnuse kõigi väärtustega, vaid tema väärtus sõltub ainult variatsioonrea keskkohas oleva väärtuse või väärtuste paari suurusest. Mediaani saame, kui viskame variatsioonreast välja nii suurima kui ka vähima väärtuse ja kordame seda protseduuri nii kaua, kuni jõuame keskeloleva väärtuseni või väärtuste paarini. Seetõttu kasutataksegi siis, kui variatsioonreas on üks või mitu väärtust, mis on ülejäänud väärtustest palju suuremad või palju väiksemad, keskväärtuse asemel mediaani.

Mood on tunnuse kõige sagedamini esinev väärtus.

Moodi tähistatakse *Mo*. Kui tunnuse väärtused on esitatud variatsioonreana, siis ei ole moodi leidmine raske, näiteks veltveebel Mesipuu koostatud saapanumbrite variatsioonreas on moodiks rühmas kõige levinum saapanumber, s.o. 43.

46 46 45 45 44 44 44 44 44 44 44 44 43 43 43 43
43 43 43 43 43 43 43 42 42 42 42 42 42 42 42 41
41 41 41 41 41 40 40 40

Võib juhtuda, et uuritav tunnus on *multimodaalne* (kahe moodi korral *bimodaalne*), s.t. et mitmel tunnuse väärtusel on võrdne, ülejäänutest suurem esinemissagedus. Näiteks variatsioonreas, milles olid VI^B klassi poiste pikkused on kaks erinevat moodi, 152 cm ja 153 cm.

148, 149, 150, 151, $\boxed{152}$, $\boxed{152}$, $\boxed{153}$, $\boxed{153}$, 154, 188.

Kui ühel tunnusel on üle kahe või kolme moodi, siis öeldakse, et selle tunnusel mood puudub.

Mood tähendab antud tunnuse kõige tüüpilisemat väärtust. Tihti käsitletakse moodi kui normi[1], näiteks normaalne abiellumisiga, normaalne palk, normaalne vastsündinu pikkus ja kaal.

Mood on ainus keskmine, mida saab kasutada nominaaltunnuste puhul.

[1] näiteks prantsuse keeles tähistab moodi sõna *normale.*

Keskmiste kasutamisest

Kui meid huvitab kõige tüüpilisem väärtus, siis seda näitab kõige suurema sagedusega väärtus mood. Mood on sageduse kõrgpunkt, ta ei näita, kas ja kui palju on temast suuremaid ja vähemaid väärtusi. Nominaaltunnuste korral (näiteks rahvus, elukutse) leitakse keskmisena mood.

Mediaani leidmisel ei arvestata tunnuse väärtusi vaid ainult suurusjärjestust. Mediaani kasutatakse siis, kui on eesmärgiks leida täpne andmete jaotuse keskpunkt, või kui andmete hulgas on ekstremaalseid väärtusi, mis oluliselt mõjutavad keskväärtust.

Keskväärtus sõltub kõigist tunnuse väärtustest, kuid ta ei pruugi ise olla tunnuse väärtus. Keskväärtus võib sattuda vahemikku, kus tunnusel on vähe väärtusi või need puuduvad hoopis. Siiski kasutatakse keskväärtust küllalt sageli, sest ta on aluseks teiste statistiliste näitajate (näiteks standardhälve, korrelatsioonikordaja) määramisele. Nende leidmisega tutvume edaspidi.

189. Kadri Kaval seletas oma pinginaabrile moodi nii: "Vaata aknast välja. Kui kõige sagedasem naiste peakate on valge barett, siis järelikult valge barett on *moes,* seega naiste peakatte mood on valge barett". Kas *mood* ja mood on omavahel seotud?

190. Kui ajaleht "Õitsev Eestimaa" väidab et Eesti keskmine palk 1996. aasta oktoobris on 3120 krooni, kas see tähendab et
a) kõigi palkade keskväärtus on 3120 krooni;
b) kõige levinum kuupalk on 3120 krooni;
c) kuupalga mõttes "keskmine" eestlane saab 3120 krooni kuus palka?
Kuidas ja kas mõjutaks keskmist palka metalli- ja puidukuningate, mõnede pankurite, parlamendisaadikute ja kõrgete valitsusametnike äkiline emigreerumine Eestist?

191. Miks on parlamendisaadikutele kasulik ja rahvale kahjulik, et rahvaesindajate keskmine palk on seatud sõltuvusse keskmisest palgast aga mitte miinimumpalgast? Et parlamendisaadikud ja kõrged riigiametnikud on osa rahvast, siis nende palk sõltub rahva keskmisest palgast. Seega parlamendisaadikute palk sõltub parlamendisaadikute palgast. Kas selline parlamendisaadiku palga määramine on matemaatiliselt korrektne?

192. Milliste tunnuste puhul lk. 54 olevast sisseastumiskatsete protokolli tabelist saab leida ainult moodi? Milliste korral keskväärtuse, mediaani ja moodi?

193. Leia oma klassi õpilaste *pikkuse* keskväärtus, mediaan ja mood. Leia need suurused ka poiste ja tüdrukute korral eraldi. Võrdle poiste ja tüdrukute karakteristikud; keskväärtusi, mediaane ja moode omavahel.

194. Leia oma klassi õpilaste *sünnikuu* ja *sünnipäeva* moodid.

195. Leia oma klassi populaarseim *lemmikloom*.

196. Leia sõdurite pikkuse sagedustabeli alusel (vt. lk. 63 ja 66) tunnuse *pikkus* mood ja mediaan.

197. Leia veltveebel Mesipuu ja referent Kaelapuu sagedustabelitest (lk. 61 ja 62) tunnuse *saapa number* mediaanid, moodid ja keskväärtused.

HAJUVUSMÕÕDUD

Näide 1. Kontrolltööd kirjutas 50 poissi ja 50 tüdrukut. Töö eest võis saada 0-80 punkti. Poiste tulemuste keskväärtus oli 55 punkti ja ka tüdrukute tulemuste keskväärtus oli 55 punkti.
Kõige nõrgema poisi tööd hinnati 8 punktiga, kõige tublimat tööd 80 punktiga. Kõigi tüdrukute tööde tulemused olid vahemikus 40 - 65 punkti. Seega võib öelda, et tüdrukute tase oli *ühtlasem* kui poiste tase.
Tunnuse iseloomustamine ainult keskmiste abil annab liiga vähe informatsiooni. Peab kasutama ka karakteristikuid, mis näitavad, kui palju keskmiselt erineb tunnuse väärtus keskväärtusest või mediaanist. Sellised karakteristikud on *hajuvusmõõdud*.

Hajuvusmõõdud iseloomustavad tunnuse väärtuste hajuvust (ehk teisiti öeldes, kas väärtused erinevad üksteisest palju või mitte).

Enimkasutatavad hajuvusmõõdud on:

1) minimaalne ja maksimaalne element;

2) variatsioonrea ulatus;

3) alumine kvartiil ja ülemine kvartiil;

4) dispersioon ja standardhälve;

5) variatsioonikordaja.

Minimaalne ja maksimaalne element

Tunnuste hajuvuse uurimisel on kõige lihtsam leida maksimaalset ja minimaalset elementi.

Minimaalne element on tunnuse väärtuste hulgas vähim ja **maksimaalne element** suurim väärtus. Minimaalset ja maksimaalset elementi tähistatakse vastavalt *Min* ja *Max*. Kõik ülejäänud tunnuse väärtused jäävad nende väärtuste vahele.

Mida suurem on maksimaalse ja minimaalse elemendi vaheline erinevus, seda suurem on tavaliselt ka tunnuse hajuvus. Seda maksimaalse ja minimaalse elemendi vahet nimetatakse **variatsioonrea ulatuseks.**

Suurema andmekogumi vaatlemisel ei pruugi andmete sisestamisel või mõõtmisel tehtud vead silma paista. Kui nüüd leida arvtunnuste minimaalsed ja maksimaalsed elemendid, siis torkab kohe silma, et Tiina Täägi pikkus on ainult 18 cm ja Miina Miini kaal 8,05 kg. Seega minimaalse ja maksimaalse elemendi leidmisest on kasu ka vigade kõrvaldamisel andmestikust.

Minimaalse ja maksimaalse elemendi leidmisest on kasu ka siis, kui tahame leida antud üldkogumi seisukohalt ebatüüpilisi objekte ja neid vaatluse alt välja jätta.

Näide 2. VI^B klassis on viisteist tüdrukut. Nende pikkuse mõõtmisel ja kaalumisel saadi järgmised tulemused:

Pikkus: 135, 152, 153, 154, 155, 155, 155, 156, 157, 158, 159, 160,162, 163, 164.

Kaal: 30, 44, 45, 45, 45, 46, 47, 49, 50, 51, 52, 52, 54, 56, 58.

Täiendaval uurimisel selgus, et kõige kergem oli ühtlasi ka lühim. Selleks osutus väike Liisi Pöial, kes oli klassiõdedest paar aastat noorem.

Minimaalse ja maksimaalse elemendi kasutamist hajuvuse iseloomustamisel takistab see, et valimi mahu suurendamisel kipub minimaalne element vähenema ja maksimaalne element suurenema. Seetõttu läheb vaja ka teisi hajuvusmõõte.

Alumine ja ülemine kvartiil

Alumine kvartiil[1] on tunnuse väärtus, millest väiksemaid (või võrdseid) liikmeid on variatsioonreas $\frac{1}{4}$ ehk 25 %.

Ülemine kvartiil on tunnuse väärtus, millest suuremaid (või võrdseid) liikmeid on variatsioonreas $\frac{1}{4}$ ehk 25 %.

Alumist kvartiili tähistatakse $\underline{Kv}$ ning ülemist kvartiili $\overline{Kv}$.

Kvartiilid on variatsioonrea alumise ja ülemise poole mediaanid. Alumise ja ülemise kvartiili vahele jäävad pooled tunnuse väärtustest. Kvartiilide erinevus näitab samuti tunnuse hajuvust. Mida suurem on kvartiilide vahe, seda suurem on tunnuse hajuvus.

Näide 3. Vaatleme veelkord VI^B klassi tüdrukute pikkuste variatsioonrida:

[1] inglise k *quarter* - veerand; ladina k *quarta pars* - neljandik.

135, 152, 153, 154, 155, 155, 155, 156, 157, 158, 159, 160, 162, 163, 164.

Selle rea mediaan on 156 cm.

135, 152, 153, 154, 155, 155, 155, $\underset{Me}{\boxed{156}}$, 157, 158, 159, 160, 162, 163, 164.

Variatsioonrea alumises ja ülemises pooles on mõlemas 7 objekti. Seitmest objektist keskmiseks on neljas objekt. Seega alumiseks kvartiiliks on 154 ja ülemiseks kvartiiliks 160.

135, 152, 153, $\underset{\underline{Kv}}{\boxed{154}}$, 155, 155, 155, $\underset{Me}{\boxed{156}}$, 157, 158, 159, $\underset{\overline{Kv}}{\boxed{160}}$, 162, 163, 164.

Ülemise ja alumise kvartiili vahe on 160-154 = 6 cm.

Lisaks kvartiilidele kasutatakse statistikas vahel ka detsiile.

Detsiilide abil jaotatakse variatsioonrida kümneks osaks.

Näiteks *esimene detsiil* on tunnuse väärtus, millest väiksemaid (või võrdseid) liikmeid on variatsioonreas $\frac{1}{10}$ ehk 10 %.

Järgnevas tabelis on toodud erineva tulutasemega perede rahalised sissetulekud ja väljaminekud 1994 aasta viimases kvartalis ühe kuu keskmisena[2].

Detsiil	I	II	III	IV	V	VI	VII	VIII	IX	X
Sissetulekud	331	541	654	781	926	1097	1335	1652	2134	3909
Väljaminekud	476	627	724	847	927	1097	1253	1586	1945	3574
Toiduained	208	270	277	302	310	348	380	430	483	608
Söömine väljaspool kodu	7	10	16	13	23	24	25	40	40	77
Alkohol	6	8	10	10	13	16	17	23	27	40
Tubakas	7	7	7	9	10	8	13	14	12	19
Keskmiselt lapsi	1,63	0,96	1,05	0,99	1,11	1,08	0,94	0,92	0,78	0,59

Kui suur osa Eesti peredest elas 1994. aasta valitud kuul varasemate säästude arvel? Kas jõukus ja kasinus käivad koos? Kes võivad lubada endale süüa sööklas ja restoranis? *Milliseid järeldusi oskad veel selle tabeli põhjal teha*?

[2] Tabelis olevad andmed on võetud 1995.aasta Statistika Aastaraamatust.