

STATISTIKA

1. STATISTIKA OLEMUS

Mõistel **statistika** on mitu omavahel seotud, kuid sisu poolest erinevat tähendust:

1. Sageli mõeldakse statistika all iseseisvat teadust, mille põhiülesandeks on meetodite väljatöötamine statistilise informatsiooni hankimiseks, töötlemiseks ja analüüsimiseks (järeldeste ja üldistuste tegemiseks) ning väljatöötatud meetodite kasutamine ülesannete praktilisel lahendamisel. Siit tulenevalt võime statistika jaotada kaheks tegevuslikult erinevaks osaks - **teoreetiline statistika** (millest põhiosa moodustab matemaatiline statistika) ja **rakenduslik statistika**. Järgneva teksti kirjutamisel on silmas peetud eelkõige rakenduslikke eesmärke.
2. Teiseks kasutatakse mõistet statistika statistilise informatsiooni kogumite tähenduses, tähistades hangitud andmete hulka.
3. Kolmandaks kasutatakse mõistet statistika ka praktilise tegevusala tähenduses. Sellisel juhul mõeldakse termini statistika all ühiskonna institutsioonilise infrastruktuuri elementi, mis tegeleb mitmesuguste, eelkõige arvuliste, andmete hankimise ja töötlemisega. Statistikaga tegelejate ring on väga lai, sisaldades eneses harilikult riiklikku statistika keskasutust (Eesti Statistikaamet), spetsialiseeritud riiklikke institutsioone või nende allüksusi, eraalgatusel tegutsevaid ettevõtteid, ühiskondlike organisatsioonide töögrupe, üksikisikuid jt.

Mõiste **statistika** sisulise poole avamiseks loetleme olulisemad probleemide grupid, millega statistika tegeleb:

- Esimene tegevusvaldkond on seotud statistilise informatsiooni hankimisega, mida sõltumata andmete hankimise viisist nimetame **statistiliseks vaatluseks**.
- Teine tegevusvaldkond on statistilise informatsiooni kompaktne ja ülevaatlik esitamine. See on **kirjeldava statistika** põhiülesandeks.
- Kolmas tegevusvaldkond on seotud reaalsuses sageli esineva, majanduses eelkõige tulevikuga seonduva, ebakindluse kirjeldamisega. Abiks selliste probleemide lahendamisel on **tõenäosusteooria**.
- Neljandaks tegevusvaldkonnaks on järeldeste tegemine uuritava objekti kohta vaatluse abil hangitud andmete põhjal. Sellega tegeleb **järeldav statistika**.
- Viies oluline probleemide ring on seotud **prognoosimisega**.
- Kuues tegevusvaldkond seondub otsustuste tegemisega ebakindlas keskkonnas mittetäieliku informatsiooni tingimustes. Sellega tegeleb **statistiline otsustuste teooria**.

2. PÕHIMÕISTED

Statistika (*statistics*) on teadus informatsiooni kogumisest, esitamisest, organiseerimisest, analüüsimisest ja kokkuvõtust, nii et andmed oleksid kergesti tõlgendatavad. Statistika jaguneb oma olemuselt kirjeldavaks statistikaks ja järeldavaks statistikaks.

Kirjeldav statistika (*descriptive statistics*) tegeleb andmete esitamise ja organiseerimisega. Kirjeldavateks statistikuteks on keskväärtused, mediaan, mood, variatsiooninäitavud jt. Seda tüüpi statistikud summeerivad numbrilist informatsiooni.

Järeldav statistika (*inferential statistics*) kasutab osakogumi ehk valimi andmeid selleks, et teha järeldusi ja üldistusi temast palju suurema kogumi kohta.

2.1 ÜLDKOGUM

Statistilist uuringut planeerides alustatakse **uurimisobjekti** määratlemisest. See defineeritakse vastavalt uurimisülesandele kui nähtus või protsess, mille kohta soovitakse teha järeldusi. Statistika uurimisobjektiks on **massnähtused** - s.o. suurest hulgast vähemalt mõningaid ühiseid omadusi või tunnuseid omavatest nähtustest koosnevad nähtused. Nähtusteks võivad olla eluta- või eluslooduse objektid, tehislikud objektid ja esemed, inimühiskonna erineva taseme ja iseloomuga komponendid, sündmused, protsessid ning mitmesugused ideaalsed objektid.

Definitsioon 2.1: Juhul kui statistiline kogum sisaldab kõiki massnähtusesse kuuluvaid üksiknähtusi või andmeid kõikide nende üksiknähtuste kohta, nimetatakse teda **üldkogumiks**.

Näide: Üldkogum võib olla näiteks terve keskkooli klass, Eesti elanikkond. Euroopa riigid vms.

Üldkogumi määratlemisel fikseeritakse mõõtmishetk ja piiritletakse see ka ruumis. Ajas ja ruumis piiritletud üldkogumi korral on sellesse kuuluvate objektide hulk tavaliselt lõplik. Lõpliku üldkogumi korral tähistatakse mõõdetavate punktide arvu e. **üldkogumi mahtu** tähega N .

Kõikne e. lausvaatlus (*census*) hõlmab objektiks oleva kogumi kõik üksikliikmed. Oma iseloomult ja taotlustelt on kõiksele vaatlusele väga lähedane rahvaloendus, mille käigus püütakse uurimisse haarata kõik inimesed mingi riigi territooriumilt.

Kõikseid järeldusi võimaldavad teha ka täielikud, kogu populatsiooni haaravad registrid. Suur osa ametlikust statistikast on kõikne, s.t. andmed kogutakse kõigi uurimisobjekti punktide kohta (kõik koolid, kõik raviasutused jm.).

Kõikse uuringuga on seotud alljärgnevad probleemid:

- objekti täpne (kontseptuaalne) piiritlemine;
- üldkogumi punktide hulga võimalikult täielik haaramine, s.t. puuduvate punktide e. puuduvate andmete arvu minimiseerimine.

2.2. VALIKUURING

Definitsioon 2.2: Kui statistilisse kogumisse kuulub ainult osa üldkogumi elementidest või osa andmeid nende kohta, siis nimetatakse seda **väljavõtukogumiks e. valimiks**.

Valimiks võib olla ükskõik milline üldkogumi alamosa.

Mittekõikne ehk osavaatlus ehk valimvaatlus hõlmab ainult osa objektiks olevast kogumist. Põhjusel, et kõikne vaatlus ei ole alati füüsiliselt võimalik, samuti majanduslikest kaalutlustest lähtudes, kasutatakse peamiselt mittekõikset vaatlust ning vaatluse tulemused üldistatakse kogu üldkogumile.

Miks me ei võta kõikset valimit kogu aeg, kui vajame informatsiooni?

- Esimeseks põhjuseks kõikse valimi võtmise juures on selle kulukus. Kuna valim on üldkogumist tavaliselt palju väiksem, siis kulutused on valimi puhul märkimisväärselt madalamad.
- Teiseks eeliseks on valimvaatluse täpsus. Kuna vähem inimesi on lülitatud valimisse, siis on intervjuerijatel võimalik iga vastaja peale kulutada rohkem aega. Lisaks väheneb küsitluse läbiviimisel vajatavate intervjuerijate arv, mistõttu on võimalik palgata paremini kvalifitseeritud uurijaid ning kulutada rohkem nende väljaõppele. See omakorda peaks kaasa tooma kogu uurimuse kvaliteedi tõusu.
- Kolmas probleem kõikse valimi võtmisel on ajakulu. Juhtudel, kus populatsioon on väga suur ja kogu informatsiooni kogumine võtaks kaua aega, on valimi kasutamine aja kokkuhoiu suhtes eelistatum.

See ei ole absoluutne, et valim on alati parem kui kõikne valim. Kõikne valim on kohasem siis, kui üldkogum on suhteliselt väike.

Valikuuringu kavandamisel määratletakse uurimisobjekt (üldkogum) ja kirjeldatakse selle objektide hulk. Mõõdetakse osa üldkogumisse kuuluvatest objektidest. Valim moodustatakse **valikueeskirja** (disaini) alusel. Valimi põhjal tehakse järeldusi üldkogumi kohta. Valimit iseloomustab **valimi maht** - n (mõõdetavate objektide arv).

2.3. STATISTILINE KOGUM

Niihästi üldkogum (kõikse uuringu puhul) kui valim (valikuuringu korral) moodustab teatavate objektide hulga, nn. **statistilise kogumi**, mida lähemalt uuritakse.

Uurimisobjektiks, mille kohta järeldusi soovitakse teha on alati üldkogum:

1. **kõikse uuringu** puhul saadakse mõõtmistulemustest teave vahetult uuritava objekti kohta;
2. **valikuuringu** puhul tuleb valimi mõõtmistulemustest teha järeldusi üldkogumi kohta, s.t. lisandub üldistamisprotsess, mille puhul tuleb arvestada juhusliku vea võimalusega.

2.4. TUNNUS. TUNNUSE TÜÜBID.

Kõik nähtused on iseloomustatavad neile omaste tunnuste abil.

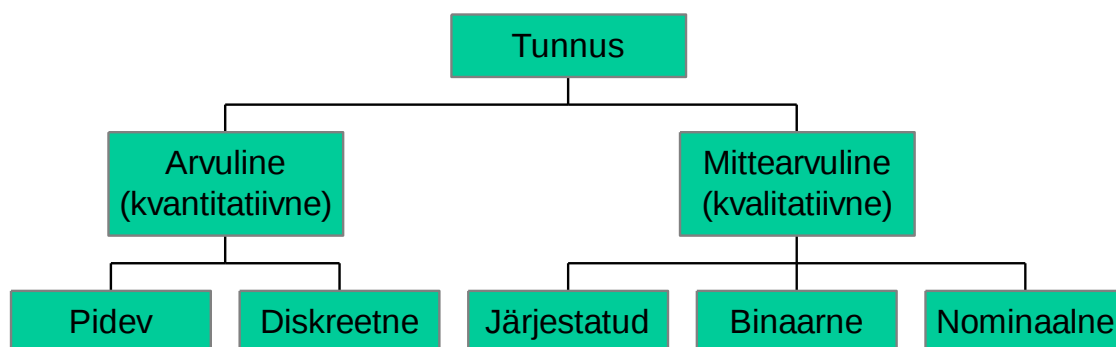
Definitsioon 2.3: Tunnus on iseloomulik omadus, mille poolest nähtused üksteisega sarnanevad või üksteisest erinevad.

Definitsioon 2.4: Tunnuseks nimetatakse üldkogumisse kuuluvate objektide mingit (arvulist või kvalitatiivselt kirjeldatavat) näitajat, mida on põhimõtteliselt võimalik mõõta. Mõõtmise võib olla kas füüsikaline (mingite füüsikaliste instrumentide, näiteks termomeetri abil) või psühholoogiline (mingi testi abil). Mõõtmiseks võib olla vaatlus, küsitlus, loendamine, mingi näidu leidmine dokumendist jmt.

Mõõtmise protsessis omistatakse igale mõõdetavale objektile mõõdetava **tunnuse väärtus**. Üldiselt eeldatakse, et tunnuse väärtus ei ole üldkogumis konstantne, s.t. et üldkogumis leidub objekte, mille puhul tunnuse väärtused on erinevad. Võimalike tunnuste väärtuste hulka nimetatakse **tunnuse väärtusvaruks**.

Sõltuvalt tunnuse väärtuse iseloomust jaotuvad tunnused:

1. arvulisteks e. kvantitatiivseteks
2. mittearvulisteks e. kvalitatiivseteks.



2.4.1. KVANTITATIIVSED TUNNUSED

Definitsioon 2.5: Kvantitatiivne tunnus - numbriliselt väljendatud tulemus.

Arvulised tunnused jaotatakse omakorda pidevateks või diskreetseteks.

1. **Pidev tunnus** võib omandada kõiki reaalarvulisi väärtusi mingist piirkonnast ehk mõõtmise käigus võidakse saada kõiki väärtusi, mis on seotud pideva skaalaga. Pidev arv-tunnus saadakse enamasti füüsikalise mõõtmise tulemusena.
2. **Diskreetsed tunnused** on tunnused, mis saavad omandada väärtusi ainult kindlate vahemike järel ehk piiratud väärtusvaruga näitajad (laste arv perekonnas). Need leitakse tavaliselt loendamise teel.

Tunnuste jaotumine pidevateks ja diskreetseteks on mõnel juhul tinglik, sest sisuliselt pidev tunnus võib osutuda diskreetseks väikese mõõtmistäpsuse tõttu. Andmete töötlemisel loetakse diskreetseks enamasti tunnus, millel on vähe väärtusi (mitte enam kui paarkümmend) ja need on täisarvud.

2.4.2. KVALITATIIVSED TUNNUSED

Definitsioon 2.6: Kvalitatiivne tunnus - mitterumbriliselt väljendatud tulemused.

Mittearvulised tunnused jaotatakse:

1. **Järjestustunnus** (enamasti psühholoogilise mõõtmise tulemus). Tavaliselt on järjestustunnuse puhul vastusevariandid ette antud, kusjuures need variandid on sisuliselt järjestatavad.
2. **Nominaaltunnus**, s.o. mittearvuline tunnus, mille vastusevariantide jaoks ei leidu sisulist (täielikku) järjestust.
3. **Kahe väärtusega (binaarne e. dihhotoomne)** tunnus, s.o. tunnus, millel on ainult kaks võimalikku väärtust (sugu). Kahe väärtusega tunnus kuulub erijuhuna järjestustunnuste hulka, sest tema väärtused on alati sisuliselt järjestatavad. Et aga binaarsete tunnuste jaoks on välja töötatud rida erimeetodeid, on sobiv neid vaadelda eraldi tüübina. Mõnikord käsitletakse dihhotoomsetena ka neid tunnuseid, millel on rohkem vastusevariante, kuid uuritava ülesande seisukohalt lähtudes võib osa variante ühendada.

2.5. TUNNUSTE KODEERIMINE

Mittearvuliste tunnuste puhul on sageli otstarbekas nad enne töötlemise algust kodeerida, st asendada sõnalised vastusevariandid arvude ehk koodidega. Kodeerimise puhul on kasulik silmas pidada alljärgnevaid otstarbekuse reegleid.

1. **järjestustunnuse kodeerimisel** tuleb jälgida, et koodid säilitaksid väärtuste sisulise järjestuse.

Seega ei tohiks kasutada kodeeringut:

hea	1
paha	2
ei oska öelda	3

Küll aga sobiksid kodeeringud:

hea	1	hea	3
ei oska öelda	2	ei oska öelda	2
paha	3	paha	1

Loomulikult võib töötlustulemuste interpreteerimisel kasutada uurijale kõige sobivamat skaalat, s.o. tunnuse väärtuste või koodide hulka. Tuleb aga jälgida, et teisendamisel säiliks tunnuse olulised omadused

2. **Binaarse tunnuse** kodeerimisel on samuti eelistatav lihtsaim võimalus, näiteks 1 ja 2 (või ka 0 ja 1, kui see on sisuliselt mõistetavam).
3. **Nominaaltunnuseid** ei ole enamasti vaja arviliseks kodeerida. Sageli on aga otstarbekas asendada pikad vastusevariandid kokkuleppeliste lühenditega.

Enne tunnuse analüüsimist tuleb kindlaks teha selle tüüp, sest erinevat tüüpi tunnustega on lubatud erinevad statistikaprotseduurid.

3. ANDMETE KOGUMINE JA ESITAMINE

3.1. STATISTILISE UURIMISTÖÖ ETAPID

Mistahes statistilisel uurimisel läbitakse harilikult kolm erinevat tööetappi:

- **statistiline vaatlus**, mille käigus hangitakse andmed uuritava massnähtuse kohta;
- **vaatlusandmete kokkuvõtt** e. esialgne töötlemine, mille jooksul minnakse üle vaadeldava kogumi üksikliikmete kohta hangitud andmetelt tervikkogumit iseloomustavatele andmetele;
- **andmete analüüs, üldistuste ja järelduste tegemine**, mille käigus toimub andmete süvendatud uurimine, nähtustevaheliste seoste otsimine ja mitmesuguste tegurite toime määramine, ning uuritava massnähtuse kohta järelduste tegemine vaatlusel kogutud andmetele tuginedes.

Konkreetses uuringu kavandamist tuleb alustada üldkogumi määratlemisest: uuringu läbiviija peab selgeks tegema, milline on teda huvitav objektide hulk. Vastavalt sellele määratlusele saab otsustada, kas on võimalik läbi viia kõikne uuring või valimvaatlus. Kui otsustatakse valimi kasuks, siis tuleb valida sobiv valimidisain. Objektiivse informatsiooni saamiseks peab valim olema juhuslik - igal üldkogumi objektil peab olema võrdne võimalus valimisse sattumiseks.

Objekti tunnuse väärtust võib registreerida (mõõta) mitut moodi: märkida üles arvuna või kirjeldada sõnaliselt. Korrektses andmestiku saamiseks peab katse kavandamisel otsustama, kuidas registreeritakse tunnuse väärtused. Statistikas öeldakse - tuleb valida skaala, millel tunnuse väärtusi mõõdetakse. Vastavalt skaala valikule tekivad erinevad tunnuse tüübid: arvilised või mitteamvilised, pidevad või diskreetsed.

3.2. ANDMETE KOGUMINE

Pärast uurimisprobleemi sõnastamist ja sobiva statistilise meetoodika valikut, peab uurija koguma andmeid, mida ta vajab analüüsi teostamiseks.

Definitsioon 3.1: Statistiliseks vaatluseks nimetatakse andmete (statistilise kogumi) hankimist uuritava nähtuse kohta. On olemas kahte tüüpi andmete kogumise allikaid: esmased ja teisesed allikad.

Definitsioon 3.2: Esmasallikate andmed ja esmased andmed on andmed, mis on kogutud spetsiaalselt uurimisel oleva probleemi või küsimuse kohta. Esmased andmed võivad olla kogutud näiteks individuaalse uuringu käigus või posti teel saadetud küsimustike kaudu. Esmased andmed on kirjastatud tegeliku andmete koguja poolt.

Näide: Statistikaamet, mis kogub esmaseid andmeid erinevate analüüside jaoks ning publitseerib saadud tulemused oma kuukirjades ja teistes väljaannetes.

3.3. ANDMETE KOGUMISE VIISID

Andmete kogumise viisidena võib ära märkida:

1. Vaatlus (*observation, survey*) otsesed e. visuaalvaatlused ning dokumentaalvaatlused;
2. Personaalne intervjuu (*personal interview*);
3. Telefoni intervjuu (*telephone interview*);
4. Iseregistreerimisega (*self-administration*) vaatlus - kui vastaja (indiviid, ettevõtte, jne.) täidab ise ankeedi või küsimustiku;
5. Registreerimine (*registration*) - on tegemist sel juhul kui vastaja on kohustatud tooma soovitud informatsiooni kindlaks määratud asukohta (näiteks autode arvele võtmine);
6. korrespondentsvaatlus;
7. ankeetvaatlus;
8. eksperiment, jt.

Teisese allika andmed ja teiseseid andmed ei olnud algselt kogutud käesoleva uurimuse jaoks, vaid mõnel muul eesmärgil.

Näide: Teisesteks andmeteks rahanduses ja arvestuses võiksid olla ettevõtete aastaaruanded, ajakirjad Wall Street Journal, või Financial Times; teiseseid allikad turunduses võivad olla mitmesugused ettevõtete müügiaruanded.

Kuigi neid andmeid erinevatest publikatsioonidest on võimalik kasutada statistilisteks analüüsideks, ei ole nad spetsiaalselt kogutud selle kindla uurimuse jaoks. Õnneks on nii äris, majanduses kui rahanduses palju häid teiseseid informatsiooni allikaid. Võimalikud probleemid teiseste allikate kasutamisega on seotud vananenud andmetega, erinevustega kasutatavates definitsioonides või andmete kogumise meetodikas ja allikate usaldusväärsusega.

Esmaste andmete peamiseks eeliseks on see, et uurija kontrollib otseselt, kuidas andmed on kogutud ning seetõttu võib ta kindel olla, et informatsioon on antud probleemi lahendamiseks oluline. Uurija võib oma

tahtmise ja vajaduse kohaselt koostada küsimustikke ja uurimusi, et saada asjakohast informatsiooni. Selle meetodi puuduseks on see, et sobiva vaatluse ja küsimustiku väljatöötamine nõuab märkimisväärselt aega, raha ning kogemusi.

3.4. ANDMETE ÕIGSUSE KONTROLLIMINE

Igasuguste andmete analüüsimise puhul on oluliseks etapiks andmete õigsuse kontrollimine. Andmetes võib esineda mitmetüübilisi vigu, loetleme neist olulisimad.

Definitsioon 3.3: Vaatlusvigadeks nimetatakse vaatlusel saadud tunnuste väärtuste hälbeid nende tõelisest väärtusest. Eristatakse metodoloogilisi vigu, registreerimisvigu ning töötlemisvigu.

1. Metodoloogilised vead on põhjustatud kasutatavate vaatlusmeetodite ja -mudelite ebatäielikkusest.
2. Registreerimisvead tekivad vaatluse käigus ja on põhjustatud vaatlusel kasutatavate vahendite ebatäiuslikkusest ning vaatelejate subjektiivsusest.
3. Olulisimateks võimalikest töötlemisvigadest on arvutusvead, aga ka ebatäpsused andmete sisestamises, jmt.

Vaatlusvigade iseloomu järgi saab vaatlusvead jaotada kaheks - juhuslikeks vigadeks ja süstemaatilisteks vigadeks.

Juhuslik viga, iseloomustab erinevusi väärtuste vahel, mis on saadud juhusliku valimi kasutamise korral ja väärtuse vahel, kui oleks uuritud kogu üldkogumit. Viga tekib juhuslikust valikust, tänu millele spetsiifilised ühikud langevad juhuslikku valimisse. Õnneks võib juhuslikku viga vähendada, kui suurendada valimi suurust ning seda on võimalik vähendada nullini, võttes uurimise alla kogu üldkogumi. Juhusliku vea suurust on võimalik matemaatilise statistika meetodite abil hinnata.

Süstemaatiline viga, on viga, mis võib tekkida nii kõikse kui valimvaatluse korral ning selle põhjuseks on halb või puudulik andmete kogumise tehnika. Kuna valim on väiksem kui kõikne valim, siis on vaja rohkem pingutada, et elimineerida mittevalimi vigu. See tähendab, et limiteeritud võimalused võivad muuta valimi palju täpsemaks kui kõikse valimi.

3.5. ANDMETE ESITAMINE

Andmete efektiivne esitus on sageli väga tähtis, sõltumata sellest, kas andmete esitamine ise on lõppeesmärk, või kasutatakse neid edaspidiseks analüüsiks ja järelduste tegemiseks. Kogutud statistiliste andmete fikseerimiseks on kolm peamist moodust.

Andmeid võidakse esitada:

1. loeteluna kirjeldavas tekstis;

2. tabelina või
3. arvjoonisena (diagrammid, graafikud jne.).

Arvandmete otsene loetlemine kirjeldavas tekstis tuleb kõne alla üksnes siis, kui refereeritavaid arve on vähe. Põhiliselt esitatakse arvandmed tabelitena. Statistilised tabelid võimaldavad esitada suhteliselt suuri andmehulki ülevaatlikult ja maksimaalselt kokkusurutud kujul. Tabeli kuju tuleb valida nii, et see väljendaks adekvaatselt uuritava nähtuse olemust ja seoseid. Tabeli kokkuleppeline kuju on järgmine:

1. iga tabeli veerg vastab tunnusele (s.t. veerg indeksiga j sisaldab tunnuse X_j mõõtmistulemusi);
2. iga tabeli rida vastab mõõdetavale objektile (s.t. i -s rida sisaldab j -nda objekti andmeid);
3. tabeli i -nda rea ja j -nda veeru lõikekohal olevas lahtris paikneb j -nda tunnuse väärtus, mis mõõdetud i -ndal objektil X_{ij} .
4. kui mingil objektil mingi tunnuse väärtus pole mõõdetud, siis on vastav koht tabelis tühi või seal paikneb puuduvat väärtust tähistav sümbol (näiteks -) ning vastav väärtus loetakse puuduvaks.

Puuduvate väärtuste tõttu on üldiselt iga tunnuse valim (s.t. objektide arv, millel selle tunnuse väärtused on mõõdetud) erinev (väiksem) valimi mahust n .

Näide: Andmete esitamine tabeli kujul

Nimi	Sugu	Vanus	Pikkus	Lapsi	Hinnang tööle
Riina	n	30	165	2	hea
Reet	n	23	175	1	olen rahul
Ain	m	45	182	3	halb
Ann	n	34	155	1	rahuldav
Alar	m	54	189	2	väga hea
Rein	m	32	179	1	ei oska öelda
Eva	n	43	167	4	hea
Hele	n	44	176	5	väga hea
Otto	m	22	188	4	hea

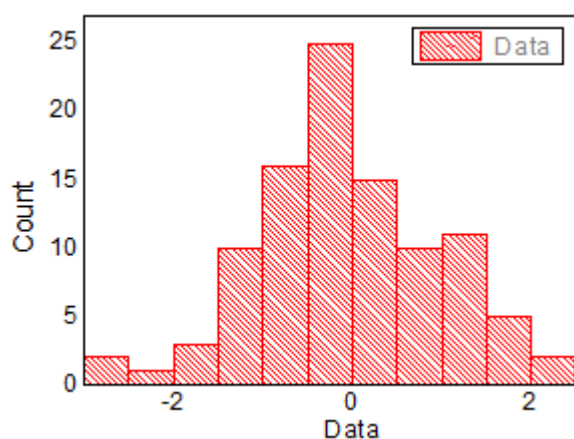
3.5.1. ARVJOONISED

Vahel on öeldud, et üks pilt on parem kui tuhat sõna ja mitte kusagil mujal ei ole see mõtteavaldus tõesem, kui andmete analüüsimisel. Tabelid on tavaliselt täidetud spetsiifiliste andmetega ja see võtab aega, et neist head ülevaadet saada. Graafikute ja diagrammide eeliseks on see, et nad esitavad andmeid rohkem ligipääsetavamal ja meeldejäävamal kujul, kuigi samas on nad vähem detailsed kui tabelid. Enamikel graafikutel ja diagrammidel on sõltumatu muutuja kujutatud horisontaalsel x -teljel ja sõltuv muutuja vertikaalsel y -teljel.

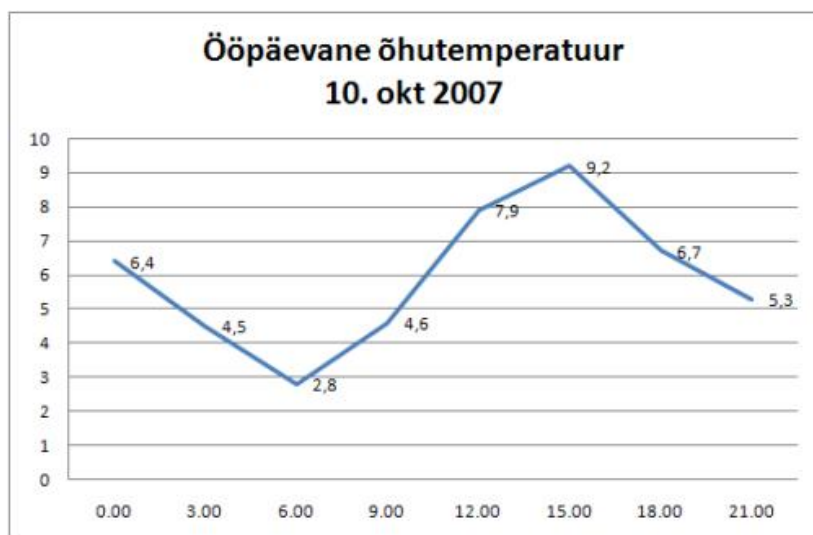
Tulpdiagrammi moodustamiseks kantakse diskreetse tunnuse korral x -teljele tunnuse erinevad väärtused, y -teljele nende väärtuste esinemissagedused. Seejärel ehitatakse x -teljele üksteisest veidi eraldatud tulbad, mille all asuv arv näitab tunnuse väärtust, tulba kõrgus aga selle väärtuse esinemissagedust. Pideva tunnuse korral on x -teljele kantud klassipiirid ja tulpdiagrammi tulpade vahet tavaliselt ei jäeta.



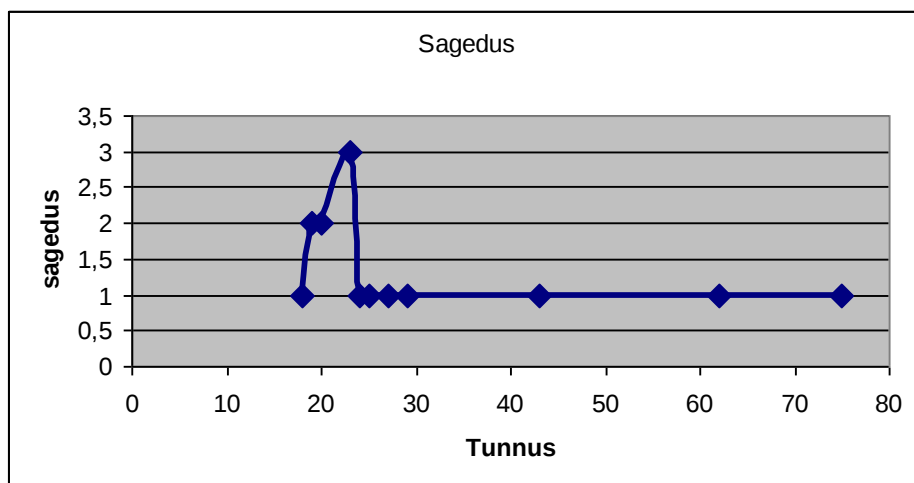
Histogramm (*histogram*) on vertikaalne tulpdiagramm sagedusjaotuse kohta. X-telg on näitaja jaoks, mida mõõdeti ja y-telg sageduse jaoks. Histogramm on kõige sagedamini kasutatav graafiku tüüp. Erinevus tulpdiagrammist seisneb selles, et kõrvuti asetsevad tulbad on üksteisega ühenduses ning ala tulba sees (kõrgus x laius) on proportsionaalne vaatluste arvule vastavas klassis.



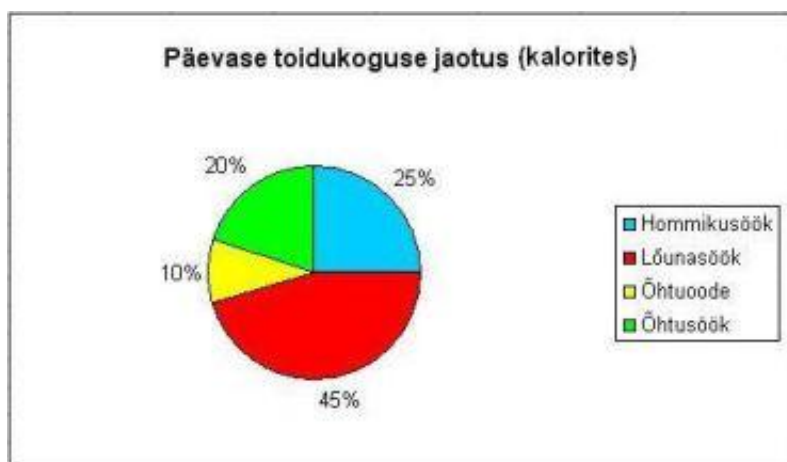
Joondigrammid (*line charts*) on konstrueeritud graafiliselt andmete punktidest ja neid ühendavatest joontest.



Sageduspolügoon (*frequency polygon*) on joongraafik, mis kirjeldab sagedusjaotust. Iga sagedus on kujutatud klassi keskpunktis. Keskpunkt on kindlaks määratud, või reaalsed klassi piirid on liidetud kokku ja jagatud kahega.



Sektordiagrammi (*pie chart*) korral vastab sektor nii suurele osale ringist, kui suure osa valimi mahust vastav klass moodustab. Kui tulpdiagrammiga võrreldakse esinevate väärtuste esinemissagedusi omavahel, siis sektordiagrammi puhul saab parema ettekujutuse terviku jagunemisest.



4. STATISTILISED READ

Vaatluse tulemusel saadakse harilikult **korrastamata statistiline rida**, milles andmed paiknevad registreerimise või laekumise järjekorras ning andmeid nimetatakse **grupeerimata andmeteks**.

Andmete paigutamisel kasvavasse või kahanevasse järjekorda, saadakse **korrastatud rida**, mida matemaatilises statistikas nimetatakse ka **variatsioonreaks**.

Nii korrastamata kui korrastatud ridu nimetatakse **lihtsateks ridadeks**. Neis on niisama palju arve kui vaatlusega hõlmatud kogumis liikmeid. Arusaadavalt pole selliste ridadega hõlpus tegelda, kui kogumid on suured. Otstarbekamad on suhteliselt lühemad read, mida on kergem töödelda ja refereerida.

5. KIRJELDAVAD ARVKARAKTERISTIKUD

5.1. SISSEJUHATUS

Definitsioon 5.1: Tunnuse jaotust ja selle omadusi iseloomustavaid suurusi nimetatakse arvkarakteristikuteks.

Tunnuse iseloomustamisel on põhilise tähtsusega kolm karakteristikute rühma:

1. **asendikarakteristikud e. keskmised**, mis annavad teavet tunnuse väärtuste paiknemise kohta arvteljel või tunnuse väärtuste hulgas. Asendikarakteristikuteks on näiteks aritmeetiline keskmine, mood, mediaan jt.
2. **hajuvuse karakteristikud e. variatsiooninäitajad**, mis iseloomustavad tunnuse väärtuse hajuvust (e. teisi sõnu - kas väärtused erinevad üksteisest palju või mitte). Hajuvuse karakteristikuteks on variatsioonirea minimaalne ja maksimaalne element, dispersioon, standardhälve, jt.
3. **jaotuse kujukarakteristikud**, mis iseloomustavad tunnuse jaotusgraafikute kuju. Kujukarakteristikutest olulisim on asümmeetriakordaja.

Nimetatud näitajad on olulised selle poolest, et nad võimaldavad kirjeldada suurt andmete hulka ainult väikese arvu kokkuvõtivate statistikutega. Kokkuvõtavad statistikud omakorda võimaldavad võrrelda individuaalseid väärtusi üldise andmekogumiga ning ka võrrelda omavahel kahte või enam andmekogumit.

Objektide arv ei kuulu kumbagi karakteristikute rühma, kuid enamasti kuulub ta karakteristikute tabeli juurde. Objektide arv annab teavet selle kohta, mitmel objektil on antud tunnuse väärtus mõõdetud. Juhul kui ühelgi objektil vaadeldava tunnuse väärtus ei puudu, siis langeb objektide arv kokku valimi mahuga. Kui aga andmestikus esineb tühikuid, on objektide arv nii mitme võrra väiksem, kui mitu mõõtmistulemust puudub.

Karakteristikute väärtusi vaadeldes tuleb arvestada seda, kas valim oli kõikne või mitte.

Definitsioon 5.2: Karakteristiku väärtust üldkogumil nimetatakse ka karakteristiku õigeks või tegelikuks väärtuseks e. parameetriks. Valimi põhjal arvutatud karakteristiku väärtust aga üldkogumi vastava karakteristiku hinnanguks e. statistikuks.

Hinnangu konkreetne väärtus on juhuslik ning sõltub konkreetsest valimist (üldkogumi objektidest võib ju moodustada palju erinevaid valimeid). Hinnangut võib uskuda juhul, kui valim on korrektselt moodustatud. Juhusliku valimi põhjal leitud hinnangu väärtus ei erine tavaliselt eriti palju üldkogumi vastava karakteristiku väärtusest ning üldiselt on erinevus seda väiksem, mida suurem on valim. Kõikse valimi karakteristikute hinnangud langevad kokku nende tõeliste väärtustega.

5.2. ASENDIKARAKTERISTIKUD

Asendikarakteristikud on kogumi ühe arvuga väljendatud koondiseloostused, mis sisaldavad eneses informatsiooni kogumisse kuuluva tunnuse väärtuste taseme kohta. Asendikarakteristikute mõõtmisel võidakse lähtuda eri vaatekohtadest, mida tingivad ühelt poolt uuritavate nähtuste iseärasused ja teiselt poolt uurimisülesanded. See tähendab, et asendikarakteristikuid võib olla palju ning et ühel kogumil võib olla mitu karakteristikut, mis formaalselt võttes on kõik üheväärselt õiged. Missugust neist kunagi vaja on, seda saab otsustada ainult uuritavate nähtuste iseloomu arvestades.

5.2.1. ARITMEETILINE KESKMINE

Aritmeetiline keskmine algandmete hulgast, mis on tähistatud $x_1, x_2, \dots, x_n$ (N näitab kogu vaatluste arvu üldkogumis) või $x_1, x_2, \dots, x_n$ (n tähistab valimi suurust), leiame me aritmeetilise keskmise, liites kõikide vaatluste väärtused ja jagades selle vaatluste arvuga. Valimi keskmist tähistatakse $\bar{x}$ ja üldkogumi aritmeetilist keskmist tähistatakse μ . Valimi aritmeetilise keskmise arvutamise valem on:

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

kus

x_i - on variantide individuaalväärtused, n - variantide arv valimis.

Üldkogumi aritmeetiline keskmine leitakse järgmise valemi järgi:

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Aritmeetilise keskmise omadused

Variantide individuaalväärtuste ja nende aritmeetilise keskmise vaheliste hälvete summa on null, s.o.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

See on aritmeetilise keskmise üks põhiomadusi. Variantide individuaalväärtuste ja nende aritmeetilise keskmise vaheliste hälvete summa on null seepärast, et positiivsete ja negatiivsete hälvete summad on võrdsed, s.t. omavahel tasakaalus.

Aritmeetilisel keskmisel on veel terve rida omadusi, mis teda iseloomustavad ja mis tema arvutamist hõlbustavad. Tänapäeval ei oma need omadused nii suurt tähtsust, kuna enamikel juhtudel töödeldakse andmeid arvutiga ja selleks mõeldud programmidega.

5.2.2. MEDIAAN

Definitsioon 5.1: Mediaan on korrastatud rea keskmine liige, millest mõlemale poole jääb võrdne arv liikmeid. Kui reas on n liiget, siis mediaani järjekorra number on: $\frac{n+1}{2}$. Kui variatsioonreas on paaritu arv liikmeid, on mediaaniks variatsioonrea keskmine liige; kui aga variatsioonreas on paarisarv liikmeid, siis asub mediaan kahe keskmise liikme vahel ning tema arvuline väärtus on võrdne nende liikmete poolsummadega.

Intervallreas leitakse mediaan valemiga:

$$M_e = x_{me} + \frac{k \left(\frac{\sum f}{2} - w \right)}{f_{Me}}$$

kus

M_e - mediaani arvuline väärtus;

x_{me} - mediaanintervalli alumine piir;

k - mediaanintervalli pikkus;

w - mediaani intervallile eelnevate intervallide sageduste summa;

f_{Me} - mediaanintervalli sagedus;

$\sum f$ - sageduste üldsumma.

Mediaanil on mõningaid eeliseid, võrreldes keskmistega. Valimi mediaan on suhteliselt lihtsalt leitav ja sobib seejuures aritmeetilise keskmise hindamiseks siis, kui tunnuse jaotus on ligilähedaselt sümmeetriline (s.t., et siis mediaan ei erine kuigi palju aritmeetilisest keskmisest). Aritmeetiline keskmine on enamikel juhtudel mingi abstraktne tunnuse väärtus, mida reas tegelikult ei esinegi. Mediaan seevastu on statistilise rea konkreetne liige.

5.2.3. MOOD

Definitsioon 5.2: Moodiks nimetatakse tunnuse kõige sagedamini esinevat väärtust.

Mood kui karakteristik pakub huvi siis, kui valimi maht on suur ja moodi esinemissagedus erineb märksa ülejäänud väärtuste esinemissagedusest. Kui tunnusel on ainult üks mood, nim. teda **unimodaalseks**. Võib juhtuda, et tunnus on **multimodaalne** (kahe moodi korral ka **bimodaalne**), s.t. et mitmel erineval tunnuse väärtusel on võrdne, ülejäänutest suurem esinemissagedus. Kui ühel tunnusel on juba rohkem kui 2-3 moodi, öeldakse ka, et sellel tunnusel mood puudub. Enamik statistikapakette väljastab mitme moodi korral ainult ühe - kas kõige väiksema või kõige suurema.

Moodi arvutamine intervallitud variatsioonreas toimub järgmise valemi abil:

$$M_o = x_{Mo} + \frac{k(f_{Mo} - f_{Mo-1})}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})},$$

kus

x_{Mo} - moodintervalli alampiir;

k - moodintervalli pikkus;

f_{Mo} - moodintervalli sagedus;

f_{Mo-1} - moodintervallile eelneva intervalli sagedus;

f_{Mo+1} - moodintervallile järgneva intervalli sagedus.

Mood kui kõige sagedasem tunnuse väärtus leiab sagedast kasutamist tarbijate vajaduste uurimisel. Moodi võib mõningal juhul vaadelda kui normi. Näiteks mood kui normaalne vastsündinu kehakaal, normaalne abiellumisiga.

5.2.4. KVANTIILID (KVARTIILID, DETSIILID, TSENTIILID)

On olemas veel rida asendikeskmisi, mis jaotavad korrastatud statistilise rea vastavaks arvuks võrdseteks osadeks liikmete arvu järgi. Taolisi asendikeskmisi nimetatakse **kvantiilideks**. Ka mediaan on kvantiil, sest ta jaotab rea kaheks võrdseks osaks liikmete arvu järgi. Kõikide kvantiilide arvutusvalemid on sarnased mediaani omale, arvestada tuleb vaid seda, kui mitmeks võrdseks osaks uuritav rida jaotatakse.

Kui jaotada rida neljaks võrdseks osaks liikmete arvu järgi, saadakse **kvartiilid**. Kvartiile on kolm: Q_1 , Q_2 , Q_3 , kusjuures $Q_2 = Me$. Esimene kvartiil on sisuliselt võrdne mediaaniga rea esimesest poolest ning kolmas kvartiil on võrdne mediaaniga rea teisest poolest. Kvartiilide arvutusvalemid intervallridade korral on järgmised:

esimene kvartiil Q_1

$$Q_1 = x_{Q_1} + \frac{k \left(\frac{\sum f}{4} - w_{Q_1} \right)}{f_{Q_1}}$$

teine kvartiil Q_2 :

$$Q_2 = x_{Q_2} + \frac{k \left(\frac{\sum f}{2} - w_{Q_2} \right)}{f_{Q_2}} = Me$$

kolmas kvartiil Q_3 :

$$Q_3 = x_{Q_3} + \frac{k \left(\frac{3 \sum f}{4} - w_{Q_3} \right)}{f_{Q_3}}$$

kus

$x_{Q_1}, x_{Q_2}, x_{Q_3}$, - vastavate kvartiiliintervallide alampiirid;

$w_{Q_1}, w_{Q_2}, w_{Q_3}$ - vastavale kvartiiliintervallile eelnevate intervallide sageduste summa;

$f_{Q_1}, f_{Q_2}, f_{Q_3}$ - vastavate kvartiiliintervallide sagedused;

Definitsioon 5.3: Asendikeskmisi, mis jaotavad korrastatud statistilise rea kümneks võrdseks osaks vaatlusandemete arvu järgi, nimetatakse detsiilideks.

Kokku on detsiile 9 ning esimene detsiil arvutatakse valemiga:

$$D_1 = x_{D_1} + \frac{k \left(\frac{\sum f}{10} - w_{D_1} \right)}{f_{D_1}}$$

Vastavaid detsiile sisaldavate intervallide leidmine toimub analoogselt mediaanile ja kvartiilidele ning teiste detsiilide arvutusvalemid tuleb vastavalt kohandada nende intervallidele.

Definitsioon 5.4: Asendikeskmisi, mis jaotavad korrastatud statistilise rea sajaks võrdseks osaks liikmete arvu järgi, nimetatakse tsentiilideks e. protsentiilideks e. pertsentiilideks.

Kokku on pertsentiile 99 ja arvutaminegi on analoogne eespool toodutele.

5.3. HAJUVUSKARAKTERISTIKUD EHK VARIATSIOONINÄITARVUD

Eelmises punktis vaatlesime erinevaid võimalusi, kuidas leida andmekogumi jaoks "tüüpilist" või "keskmist" väärtust.

Kui kogumi liikmetel ühe ja sama tunnuse arväärtused erinevad, siis tema väärtus **varieerub**. Keskmised on kogumite kvantitatiivsete üldistustena ühekülgised selles mõttes, et nad kajastavad üksnes variantide arväärtuste üldist keskmist nivood, mitte aga nende omavahelisi erinevusi. Keskmised ei heida valgust andmete hajuvusele e. varieerumisele. Praktikas kasutatakse paralleelselt keskmistega ka **variatsiooni**-

näitarve, mis iseloomustavad kogumi üksikliikmete kõrvalekaldumist kesktasemest. Kirjeldada aga andmete hajuvust mingis variatsioonreas on palju keerulisem ülesanne kui oli seda "tüüpilise" või "keskmise" väärtuse kirjeldamine. Põhjuseks on, et variatsiooninäitarvud ei ole kergesti intuiitselt tõlgendatavad ning nende sisust on vahel raske aru saada. Variatsiooninäitarvude puhul kehtib aga reegel, et kui andmed on tihedalt koondunud keskmise väärtuse ümber, siis saame suhteliselt väikesed väärtused erinevatele variatsiooninäitarvudele, kui aga andmed on hajutatud, siis on tulemuseks suhteliselt suured väärtused.

Oma olemuselt jagunevad variatsiooninäitarvud absoluutseteks ja suhtelisteks;

- **absoluutsed näitajad** on seotud mõõtmisdimensiooniga ning seetõttu võimaldavad võrrelda vaid sarnastes mõõtühikutes mõõdetud andmehulki;
- **suhtelised näitajad** arvutatakse kas osatähtsustena või protsentaalselt ning nad võimaldavad võrdlusi ka erinevates mõõtühikutes väljendatud andmekogumite puhul.

Variatsiooninäitarvude praktilised kasutamisvõimalused on üsna laialdased. Lisaks keskmiste abil tehtavate üldistuste täiendamisele, kasutatakse neid veel valikvaatluste, korrelatsiooni analüüsi ja regressioonanalüüsi raames.

Tuleb mainida, et samuti nagu keskmiste juures olid mõned näitajad teatud situatsioonides sobilikumad kui teised, siis sama kehtib ka variatsiooninäitarvude juures.

5.3.1. ABSOLUUTSED VARIATSIOONINÄITARVUD

Minimaalne ja maksimaalne element

Tunnuse hajuvuse iseloomustamisel on kõige lihtsam vaadelda valimi maksimaalset ja minimaalset elementi. **Minimaalne element** on tunnuste väärtuste hulgas vähim ja **maksimaalne element** suurim mõõdetud väärtus. Minimaalset ja maksimaalset elementi tähistatakse vastavalt X_{min} ja X_{max} . Kõik ülejäänud väärtused jäävad konkreetse valimi puhul nende kahe väärtuse vahele. Mida suurem on minimaalse ja maksimaalse elemendi erinevus, seda suurem on reeglina ka tunnuse hajuvus.

Variatsiooniamplituud (*range*)

Kõige hõlpsamalt saab ettekujutuse tunnuse väärtuse varieerumisest **variatsiooniamplituudi e. -ulatuse** vahendusel, mis leitakse rea kõige suurema ja väiksema liikme arväärtuse vahena, s.o.

$$R = X_{max} - X_{min} ,$$

Variatsiooniamplituud iseloomustab rea liikmete väärtuste erinevust, andes nende maksimaalse võimaliku erinevuse väärtuse. Variatsiooniamplituud on seda suurem, mida suuremad on rea üksikliikmete erisused. Kui kõikide variantide individuaalväärtused on võrdsed, siis $R = 0$.

Dispersioon

Dispersioon ehk **keskmise ruuthälve** on variantide individuaalväärtuste ja nende aritmeetilise keskmise vaheliste hälvete ruutkeskmise. Sisuliselt leitakse iga elemendi erinevus keskväärtusest, aga et arv oleks positiivne, tõstetakse ta ruutu. Seejärel leitakse nende erinevuste ruutude keskväärtus. Dispersiooni valem on lihtsa statistilise rea puhul

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ valim;}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \text{ üldkogum}$$

Suurus $n - 1$ on valimidispersiooni avaldises selleks, et saada hinnangut, mis ei sisalda süstemaatilist viga (nihet). Kui valimi maht on suur (100 ja rohkem), võib nimetajas olla n .

Praktilistes arvutustes on enamasti otstarbekas kasutada sama valemi teist kuju

$$s^2 = \frac{1}{n-1} \left(\sum x_i^2 - n\bar{x}^2 \right)$$

Intervallrea puhul arvutatakse dispersioon valemiga:

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}$$

Nagu näha, leitakse dispersioon üldjoontes samuti nagu keskmise lineaarhälve. Variantide individuaalväärtuste ja nende aritmeetilise keskmise vaheliste hälvete $(x_i - \bar{x})$ märkide kaotamiseks ei lülitata aga arvestustesse mitte nende absoluutväärtusi, vaid tõstetakse nad ruutu. Dispersiooni mõõtühikuks on variandi dimensiooni ruut. On σ^2 mõõtühikuks meeter, siis on dispersioon ruutmeetrites. Tihti saadakse aga niiviisi kas raskesti lahtimõtestatavad või üldse mõttetud ühikud, kui variantide mõõtühikuteks on näiteks tükk, rubla, protsent. Erinevate valimite dispersioone võib aga omavahel võrrelda küll ning kehtib lihtne reegel: mida suurem dispersioon, seda suurem hajuvus.

Standardhälve

Dispersiooni tõlgendusraskused on hõlpsasti ületatavad dispersioonist ruutjuure leidmise teel. Dispersiooni ruutjuur annab **standardhälbe** e. **keskmise ruuthälbe** s . Seda leitakse ühega järgmistest valemitest:

$$s = \sqrt{s^2}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \text{ valim}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} \text{ üldkogum}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}} \text{ intervallida}$$

Standardhälve on alati samades mõõtühikutes, milles variandidki ning sellega võib opereerida põhimõtteliselt samuti nagu keskmise lineaarhálbega. Vahemik $\bar{x} \pm \sigma$ hõlmab 68,27% normaalse jaotuskõvera alusest pindalast.

Moodi, aritmeetilise keskmise ja mediaani omavaheline sõltuvus

Mõningatel juhtudel võib moodi leidmine olla raskendatud kas andmete puudumise või siis rea erilise iseloomu tõttu, milles moodi on ligikaudsetegi meetoditega raske arvutada. Kui aritmeetiline keskmine ja mediaan on varem teada, võib mõõdukalt ebasümmeetrilistes ridades leida moodi valemiga

$$M_o = \bar{x} - 3(\bar{x} - M_e)$$

Põhimõtteliselt tähendab see valem mõõdukalt ebasümmeetrilistes ridades hästi silmahakkavat tõsiasja, et mediaan paikneb aritmeetilise keskmise ja moodi vahel, olles seejuures aritmeetilisele keskmisele lähemal (1/3 vahemaad aritmeetilisest keskmisest, 2/3 moodist).