TALLINNA
PEDAGOOGIKAÜLIKOOL

INFORMAATIKA OSAKOND

Statistika loengumaterjale

Katrin Niglas

1997

Sisukord

1. Sissejuhatus

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. Sellegipoolest on õpilased, 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 õpilased unustavad sootuks üldised ideed, mida need arvutused illustreerima peaks. Lugejatel on raske näha arvutuslike puude taga statistilist metsa.

Seepärast ei pöörata kogu järgnevas käsitluses tähelepanu mitte valemitele ühe või teise statistiku arvutamiseks vaid püütakse selgitada statistiliste ideede (kontseptsioonide) olemust sõnade, näidete ja jooniste abil.

Loengumaterjalide koostamisel on kasutatud D. Rowntree raamatut "Statistics without tears".

2. Mis on statistika?

2.1 Statistiline mõtteviis.

Statistiline mõtteviis on meile kõigile igapäevasest elust tuttav ja omane.

Võtame ühe lihtsa näite: ma ütlen teile, et ma 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: “Ma käin kinos keskmiselt kaks korda kuus” või “Sügisel on oodata palju vihma” 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 ennustus üksiku aasta või siis õpilase kohta.

Tihtipeale 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 uurimuse, mis sisaldab ANDMETE kogumist antud nähtuse kohta, kogutud andmete töötlemist ning põhjendatud järelduste tegemist.

Statistilise maailmavaate keskseks mõisteks on TÕENÄOSUS, s.t. statistika ei anna meile kunagi 100% kindlust, eriti kui tegeldakse üksiku inimese või sündmusega, vaid lubab määrata, kui suur on võimalus selle sündmuse toimumiseks.

Statistiline mõtteviis on mõistmine, et meie vaatlused (mõõtmised) ei saa kunagi olla täiesti täpsed ning, et meie oletus (hüpotees) võib kehtida näiteks 95-l (või 99-l) juhul 100-st, kuid mitte kunagi 100-l juhul 100-st.

Näiteks laps, kelle pikkuseks me oleme mõõtnud 162 cm, ei ole täpselt nii pikk - tema pikkus võib olla kuskil 161,75 cm ja 162,25 cm vahel, kuid mitte täpselt 162 cm. Ning kui me kasutame olemasolevaid vaatlusandmeid järelduste tegemiseks teiste (mitte mõõdetud) objektide kohta, siis on meil võimalus eksida veel palju suurem. Näiteks juhul, kui me tahame ennustada ühes klassis käivate laste mõõtmisel saadud keskmise pikkuse põhjal teises klassis käivate laste keskmist pikkust.

Seepärast ei saa me olla täiesti täpsed, kuid statistika võimaldab meil määrata oma vigade ulatuse.

Seega me võime peaaegu täpselt väita, et lapse pikkus on vahemikus $162 \pm 0{,}25$ cm; ning me võime arvutada, et 99-l juhul 100-st on laste keskmine pikkus teises klassis näiteks vahemikus 162 ± 3 cm.

2.2 Kirjeldav ja järeldav statistika. Üldkogum ja valim.

Enamuses statistika käsitlustes tõmmatakse selge piir kahe statistika valdkonna vahele:

1. KIRJELDAV STATISTIKA, mis pakub meetodeid (vaatlus)andmetest kokkuvõtete tegemiseks ja nende kirjeldamiseks ning
2. JÄRELDAV STATISTIKA, mis kasutab kogutud (vaatlus)andmeid baasina hinnangute ja prognooside tegemiseks (veel) mitte vaadeldud situatsioonide kohta.

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

* "Ma käin kinos keskmiselt kaks korda kuus"
* "Sügisel on oodata palju vihma"
* "Mida varem sa kordama hakkad, seda paremini sul eksamil läheb"

* * *

Esimene lause on kirjeldav, teine ja kolmas aga ei piirdu vaid kogetu kokkuvõtmisega, vaid nendes tehakse järeldus selle kohta, mis tulevikus tõenäoliselt juhtub.

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 meie poolt püstitatud järeldused, oletused või prognoosid kehtivad.

Näiteks võivad erinevad teadlased 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 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 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 teadlased ületavad kättesaadava informatsiooni piiri: nad üldistavad tulemusi valimilt üldkogumile, nähtult ja kogetult mittenähtule ja mittekogetule.

Tulles tagasi kirjeldava ja järeldava statistika mõistete juurde, võime öelda, et kirjeldav statistika tegeleb valimi (vaatlemisel saadud andmete) resümeerimise ja kirjeldamisega, järeldava statistika ülesanne on aga üldistuste tegemine laiema objektide hulga - üldkogumi - kohta.

Kui täpsed on aga sellised üldistused osalt tervikule? See ongi küsimus, millega statistika laias laastus tegeleb: ta määrab meie eksimise tõenäosuse.

2.3 Statistilised tunnused. Tunnuste tüübid.

Vastavalt sellele, mida me uurida tahame, koosneb meie valim kas üksikutest inimestest, valgetest hiirtest, kalendrikuudest, mingitest toodetest, kartulipõldudest või millest tahes. Kõiki valimisse kuuluvaid indiviide nimetatakse statistikas OBJEKTIDEKS. Kõigil ühte valimisse kuuluvatel objektidel on mingid iseloomulikud TUNNUSED, mis meid huvitavad, näiteks: värv, sugu, hind, kaal jne. Iga üksik valimi liige erineb teistest mõne tunnuse VÄÄRTUSE poolest: mõned objektidest on ühte värvi, mõned teist; mõned on naised, teised mehed; mõned on kallimad, teised odavamad jne. Statistilised tunnused on vahendiks, mis lubab meil üksikuid objekte üksteisest eristada.

Oletame näiteks, et te tahate osta kasutatud jalgratast. Millised on need tunnused, mille põhjal te oma valiku teeksite ehk, milliseid andmeid te tahaksite erinevate rataste kohta teada, et neist endale sobiv välja valida?

* * *

Toon mõned tunnused, mis oleks minu jaoks olulised. Teie nimekiri võib olla pikem või lühem, sisaldada osasid toodud tunnustest või kõiki jne:

Jalgratta tüüp (N. naiste-, meeste-, laste-, sportratas jne.)
Valmistaja riik
Värvus
Seisukord (N. hea, rahuldav, halb)
Vanus
Hind
Käikude arv

Iga üksik jalgratas, pakutavate hulgast, erineb teistest mõne tunnuse väärtuse poolest. See, kuidas me aga erinevaid jalgrattaid nende tunnuste põjal hindame, sõltub tunnuse tüübist.

Tunnusega "jalgratta tüüp" jagame me pakutavad jalgrattad *kategooriatesse* kasutades lihtsalt nende nime, N. naisterattad, lasterattad, meesterattad jne. Kõiki selliseid tunnuseid, mis liigitavad üksikud objektid mingitesse klassidesse (kategooriatesse), kasutades selleks sõnu, nimetataksegi KATEGORIAALSETEKS e KVALITATIIVSETEKS TUNNUSTEKS.

Millised tunnused ülaltoodutest on sinu arvates veel kategoriaalsed?

* * *

Täpselt! 'Valmistaja riik' ja 'värvus' on kategoriaalsed tunnused. Esimese puhul nendest on kategooriateks erinevad riigid N. Venemaa, Soome, Saksa jne. ning teise puhul jagatakse rattad klassidesse nende värvi põhjal. Selliseid tunnuseid nimetatakse tihti ka NOMINAALSETEKS TUNNUSTEKS (ladina k. *nominalis* = nimi).

Kuid samuti on tunnus "seisukord" kategoriaalne, sest ta jagab jalgrattad kolme gruppi: heas, rahuldavas ja halvas korras olevateks. Kas sa märkad erinevust kahe eelneva tunnuse ja selle tunnuse vahel?

* * *

Tõepoolest, tunnuse "seisukord" abil võime me öelda, et ühed jalgrattad on teistest paremad: seega, me võime jalgrattad selle tunnuse põhjal järjekorda seada. Kõiki selliseid tunnuseid, mille puhul me saame öelda, et üks valimi liige on teistest parem või suurem või kiirem - ühesõnaga, saame objekte järjestada, nimetatakse JÄRJESTUS- ehk ORDINAALSETEKS TUNNUSTEKS. Pane tähele, et järjestustunnuse väärtusteks võivad olla ka numbrid (näiteks võime me kümme pakutavat jalgratast panna seisukorra järgi täielikku järjekorda: 1-kõige parem, 2-järgmine, ...,10-kõige halvem), kuid siin me kasutame numbreid tähenduses: esimene, teine, kolmas jne.

Me ei saa öelda, et esimene jalgratas on täpselt kaks korda parem kui teine või kümnes täpselt kümme korda halvem kui esimene.

Teise põhilise tunnuste tüübi moodustavad kõik need tunnused, mille väärtusteks on numbrid. Siin me saame öelda, kui palju erineb iga üksik objekt teisest; me saame seda erinevust täpselt mõõta (või loendada). Millised eelpool toodud tunnustest sa paigutaksid sellesse tüüpi?

* * *

Jalgrataste "vanus", "hind" ja "käikude arv" on kirjeldatavad konkreetsete numbriliste suurustega. Me saame öelda täpselt, mitu korda on üks jalgratas teisest kallim või kui palju on üks ratas teisest vanem ning ka käikude arv erinevatel ratastel on täpselt võrreldav. Kõiki tunnuseid, mille väärtusi me saame täpselt mõõta või loendada, nimetatakse KVANTITATIIVSETEKS TUNNUSTEKS.

Kuid samuti, nagu kategoriaalsete tunnuste puhul on ka kvantitatiivseid tunnuseid kahte tüüpi: DISKREETSED ja PIDEVAD TUNNUSED. Diskreetne on tunnus, mille võimalikud väärtused on üksteisest selgelt eraldatud. Klassikaline näide sellisest tunnusest on laste arv peres: peres võib olla 1 laps või 2 last või 3 või 4 või jne.

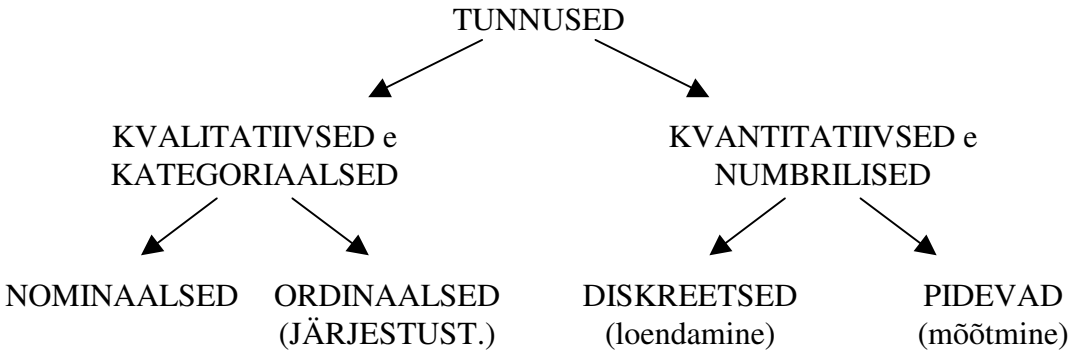
Pidevate tunnuste puhul on aga vastupidi: võttes millised tahes kaks võimalikku väärtust, võime me alati leida väärtusi nende vahel, mis on samuti võimalikud. Mäletate näidet laste pikkuse mõõtmisest? Laps võib olla praegu 149 cm pikk, kuid aasta möödudes on tema pikkus 155 cm. Kuid vahepeal pole tema pikkus olnud mitte ainult 150 cm, 151 cm, jne. vaid ka näiteks 151.5 cm, 153.3754 cm jne. Seega, laps ei kasva 1 sentimeeter või pool sentimeetrit korraga vaid tema pikkus suureneb pidevalt.

Üldiselt peame me diskreetsete tunnuste väärtuste leidmiseks kasutama loendamist ning pidevate tunnuste puhul mõõtmist. Millised meie jalgrataste tunnustest on diskreetsed ja millised pidevad?

* * *

'Käikude arv' on tõesti diskreetne tunnus. Jalgrattal võib olla, kas 1, 3, 4, 5, 8 või 10 käiku, kuid vahepealsed väärtused ei ole võimalikud. 'Vanus' on aga pidev tunnus: me võime vanust mõõta kuitahes täpselt (st. me saame alati leida vanuse, mis on näiteks 3 aasta 9 kuu ja 3 aasta 10 kuu vahel jne.). Tavaliselt tekitab vaidlusi tunnuse 'hind' paigutamine ühte või teise tunnuse tüüpi. Kui me aga mõtleme eelmiste näidete peale, siis näeme, et 'hind' on diskreetne tunnus, sest ei saa leida reaalselt võimalikku hinda näiteks 90 ja 95 sendi vahel. (NB! isegi täisarvuline hind 92 senti ei ole võimalik!) Ka eestikeelne väljend: raha lugema, näitab, et tegemist on diskreetse tunnusega. Me loeme raha, mitte ei mõõda.

Järgnev joonis illustreerib seost erinevate tunnuse tüüpide vahel:



Oluline on teada, et statistikas tuleb erinevatesse tunnuse tüüpidesse kuuluvaid andmeid käsitleda erinevalt. Kõige suurem vahe, mida tuleb andmete käsitlemisel silmas pidada, on vahe *kategoriaalsete* ja *kvantitatiivsete tunnuste* vahel.
Selle punkti lõpetuseks tahaks veel mainida, et kõiki kvantitatiivseid tunnuseid on võimalik muuta kategoriaalseteks. Näiteks võime me jagada inimesed pikkuse põhjal klassidesse: väga pikad, pikad, keskmised, lühikesed ja väga lühikesed. Nii tehes kaotame me aga informatsiooni, ning algandmete puudumisel me vastupidist teisendust (kategoriaalsest tunnusest kvantitatiivseks) teha ei saa. Selline kategoriseerimine on aga vajalik, kui me tahame erinevaid gruppe omavahel võrrelda. Gruppide moodustamist kasutatakse vahel ka selleks, et lihtsustada andmete käsitlemist.

3. Andmete kirjeldamine ehk kuidas saada kogutud andmetest paremat ülevaadet.

3.1 Tabelid ja diagrammid.

Jättes vahele andmete kogumise etapi, oletame nüüd, et teie käsutuses on hulk pabereid täis vaatlustel saadud tulemusi (ehk andmeid). Esimene asi, mis teil tuleb teha, on need andmed korrastada nii, et teie ise ning ka teised inimesed saaksid kogutud vaatlustulemustest selge ülevaate.

Võtame jällegi ühe lihtsa näite: kõrgkool viis läbi uurimuse, kus viiekümne tudengi käest küsiti muuhulgas ka seda, millist transpordi liiki ta kooli jõudmiseks kasutab.

Kõige klassikalisem viis selliste andmete korrastamiseks on koostada SAGEDUSTABEL:

Kooli jõudmiseks kasutatavad transpordivahendid

Jalgrattaga	III	3
Jalgsi	IIIII IIII	9
Mootorrattaga	II	2
Autoga	IIIII I	6
Bussiga	IIIII IIIII IIII	14
Trammiga	IIIII IIIII IIIII	16
		Kokku 50 tudengit

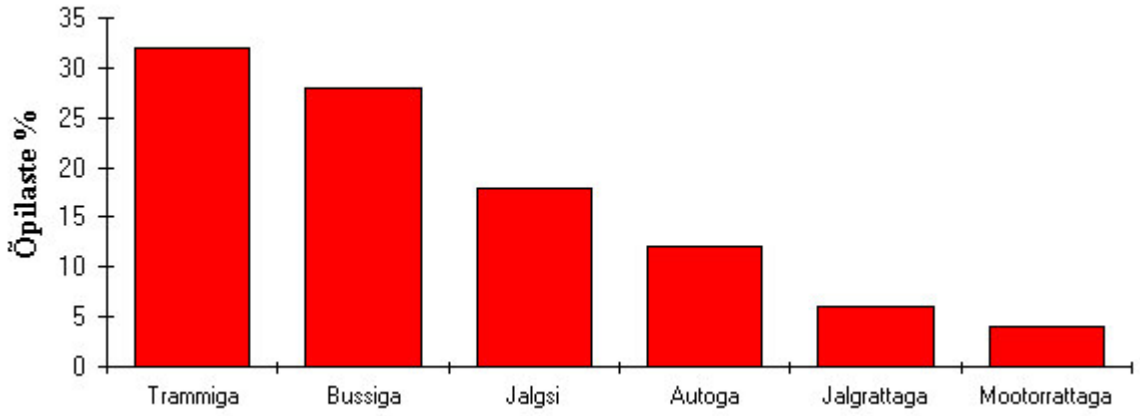
Kuid tavaliselt huvitavad meid valimi puhul mitte niivõrd ühe või teise kategooria sageduse absoluutarvud vaid proportsioonid. Seetõttu on mõistlik sagedustabel järjestada kategooriate suuruse järgi ning välja arvutada ka protsendid:

Kooli jõudmiseks kasutatavad transpordivahendid

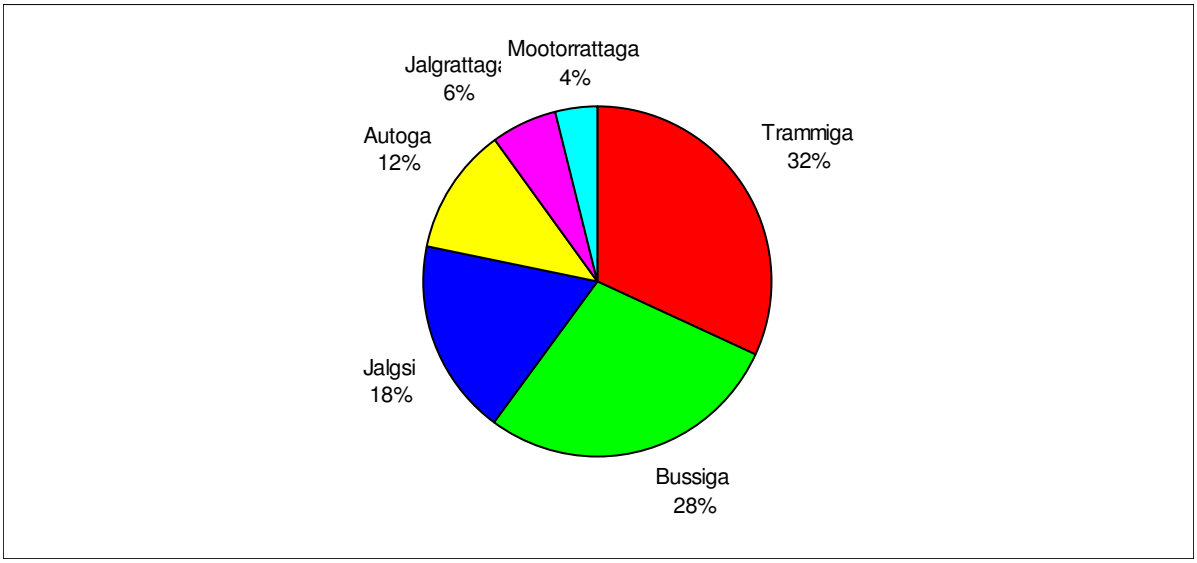
Transpordi liik	Seda liiki kasutavate õpilaste %	Seda liiki kasutavate õpilaste arv
Trammiga	32	16
Bussiga	28	14
Jalgsi	18	9
Autoga	12	6
Jalgrattaga	6	3
Mootorrattaga	4	2
Kokku	100%	50

Tänapäeval, kus andmete käsitlemisel kasutatakse üha laiemalt arvuteid, hakkavad eelpool mainitud tabelid aga tasapisi kasutusest kõrvale jääma, sest arvuti võimaldab ühe sammuga lisaks proportsioonide väljaarvutamisele koostada ka diagrammi, mis neid proportsioone illustreerib. Koostame TULPDIAGRAMMI, kus iga tulba kõrgus on proportsionaalne vastavasse kategooriasse kuuluvate õpilaste arvuga:

Kooli jõudmiseks kasutatavad transpordivahendid



Kategoriaalsete andmete proportsioonide illustreerimiseks kasutatakse ka SEKTORDIAGRAMMI. Siin on ring jagatud sektoriteks nii, et iga sektori suurus on proportsionaalne antud kategooria sagedusega.



Tulpdiagramm on ülevaatlikum juhul, kui me tahame võrrelda erinevate kategooriate sagedusi omavahel, sektordiagramm aga juhul, kui me tahame näha iga üksiku kategooria osa tervikus.

Oletame nüüd, et meid huvitab *kas*? ja *kuidas*? erinevad meeste ja naiste poolt eelistatavad kooli jõudmise meetodid. Selleks tuleks koostada nn. RISTTABEL, kus naiste ja meeste sagedused on toodud erinevates ridades:

	naised (sagedused)	mehed (sagedused)
Jalgrattaga	1	2
Jalgsi	6	3
Mootorrattaga	0	2
Autoga	2	4
Bussiga	6	8
Trammiga	10	6
Kokku	25	25

Sellised risttabeleid on võimalik koostada mistahes kahe tunnuse jaoks.

Vaatame nüüd kuidas kokku võtta numbrilisi andmeid , st. andmeid, mis kuuluvad kvantitatiivsesse tunnuse tüüpi. Meil on olemas andmed 50 õpilase pulsisageduse kohta. Toome tulemused sellises järjekorras, nagu nad mõõtmisel saadi:

50 tudengi pulsisagedused (lööki minutis)

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 pulsisagedusest? Kui kerge on leida kõige kõrgemat ja kõige madalamat pulsisagedust? Kas pulsisagedused on jagunenud ühtlaselt minimaalse ja maksimaalse väärtuse vahel või on mõned pulsisagedused tihedamini esinevad kui teised?

Neile küsimustele oleks palju lihtsam vastata, kui meie pulsisagedused oleks järjestatud suuruse järgi. Teeme seda:

50 tudengi pulsisagedused (lööki minutis)

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

Sellist rida, kus me oleme kvantitatiivse tunnuse väärtused järjestanud nende suuruse järgi nimetatakse VARIATSIOONIREAKS e.JAOTUSEKS.

Nüüd on meil lihtne leida minimaalne ja maksimaalne pulsisagedus: 62 ja 96 lööki minutis. Need väärtused võimaldavad meil lihtsalt leida jaotuse ULATUSE, milleks on maksimaalse ja minimaalse väärtuse vahe. Meil 96 miinus 62 annab ulatuseks 34 lööki minutis.

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 üliõpilase 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 keskmise väärtuse vahel. Meie näites tudengite pulsisageduste 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 lööki minutis.

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

$$\frac{0+2+3+4+6+6+10}{7} = \frac{31}{7} \approx 4.4 \text{ tundi nädalas.}$$

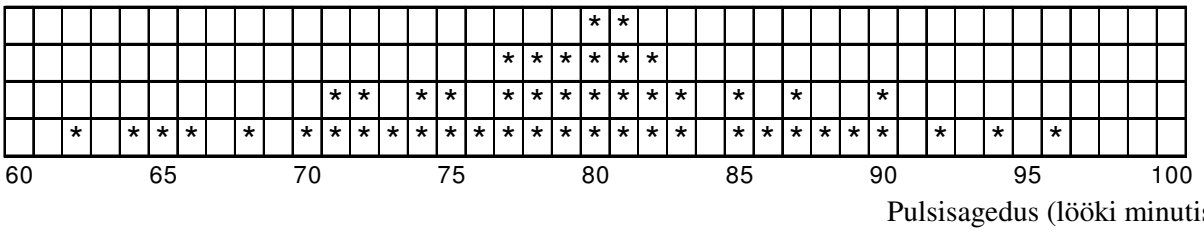
Et mitte tülitada teid 50 pulsisageduse kokkuliitmisega ning saadud summa 50-ga jagamisega, siis ütlen teile, et tudengite keskmine pulsisagedus (ehk pulsisageduste aritmeetiline keskmine) on 79.1 lööki minutis. Kui te nüüd võrdlete kahte erinevat keskmist tendentsi väljendavat suurust: mediaani ja aritmeetilist keskmist, siis te näete, et nad on natuke erinevad. Hiljem näeme, millisel juhul on kasulikum ühte või teist näitajat kasutada.

Oleme nüüd vaadanud mitut erinevat võimalust oma andmete kirjeldamiseks, kuid kas meil on praegu ettekujutus jaotuse üldisest kujust st. kas me saame seniste sammude põhjal vastata ka viimasel küsimusele, mis puudutas pulsisageduste paiknemist minimaalse ja maksimaalse väärtuse vahel?

* * *

Tõepoolest, selget pilti pulsisageduste paiknemisest variatsioonireale pealevaadates ei saa. Kui me aga koostame "punkt-diagrammi", st. märgime skaalal iga mõõdetud väärtuse punktiga, siis näeme, et palju sagedamini esinevad pulsisagedused, mis on lähedal (ulatuse) keskpunktile.

Iga punkt tähistab ühte tudengit (kokku 50)



Pulsisagedus (lööki minutis)

Andmete esitust ülaltoodud diagrammi kujul nimetatakse SAGEDUSJAOTUSEKS. See diagramm näitab, mitu korda iga väärtus mõõtmisel tulemuseks saadi, st. ta näitab iga väärtuse esinemise sagedust. Kui mitmel tudengil oli pulsisagedus 78 (90) (69) lööki minutis?

* * *

Pulsisagedus 78 on mõõdetud kolmel tudengil (sagedus=3), kahel tudengil on pulsisagedus 90 ning mitte kellelgi ei ole mõõdetud pulsisageduseks 69 lööki minutis.

Milline väärtus esineb kõige sagedamini ehk millise väärtuse esinemise sagedus on kõige suurem?

* * *

Kõige rohkem (neljal korral) on pulsisageduseks mõõdetud 80 ja 81 lööki minutis. Sellist jaotuse väärtust, mis esineb kõige sagedamini nimetatakse MOODIKS. Antud näites toodud jaotusel on seega kaks moodi: 80 ja 81 (need pulsisagedused on kõige "moodsamad" ehk kõige sagedamini esinevad).

Moodi kasutatakse kõige rohkem kategoriaalsete tunnuste iseloomustamiseks. Oletame näiteks, et 50-st küsitletust 27 olid abielus, 15 vallalised ning 8 lahutatud. Modaalne klass (ehk kategooria) on siin kahtlemata "abielus". Pange tähele, et kategoriaalsete tunnuste puhul me aritmeetilist keskmist ega tavaliselt ka mediaani arvutada ei saa!

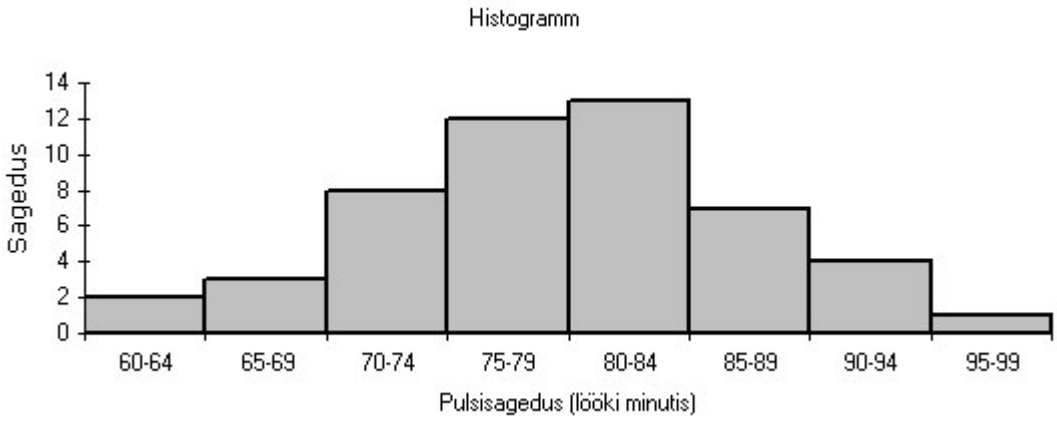
Pöördume nüüd tagasi meie pulsisageduste näite juurde. Sageli (eriti suurte andmehulkade puhul) on aga kasulik vaatlusandmed grupeerida. Näiteks võime me küsida mitu mõõtmistulemust on vahemikus 60-st 64-ni, mitu 65-st 69-ni, mitu 70-st 74-ni jne. Kui me oma andmeid niimoodi grupeerime, saame järgmise tabeli:

Pulsisagedus (lööki minutis)	Õpilaste arv (sagedus)
60-64	2
65-69	3
70-74	8
75-79	12
80-84	13
85-89	7
90-94	4
95-99	1
	Kokku 50

Sellest tabelist on jaotuse üldine kuju veelgi selgemalt näha - meie näites "kuhjuvad" vaatlusandmed jaotuse keskel. Kuid selline grupeerimine toob endaga paratamatult kaasa informatsiooni kao. Jaotuse üldise kuju selgitamisel tuuakse ohvriks üksikud väärtused.

Ülaltoodud tabeli graafiliseks esituseks on HISTOGRAMM. See on tulpdiagramm, kus iga väärtuste vahemikku tähistab ristkülik, mille kõrguseks on vastava vahemiku sagedus (või osakaal protsentides).

50 tudengi pulsisagedused (lööki minutis). HISTOGRAMM.



Ülal toodud histogrammil on grupis "70-74 lööki minutis" kaks korda rohkem liikmeid kui grupis "90-94 lööki minutis" ning seega ka tema tulp on viimase grupi omast kaks korda suurem.

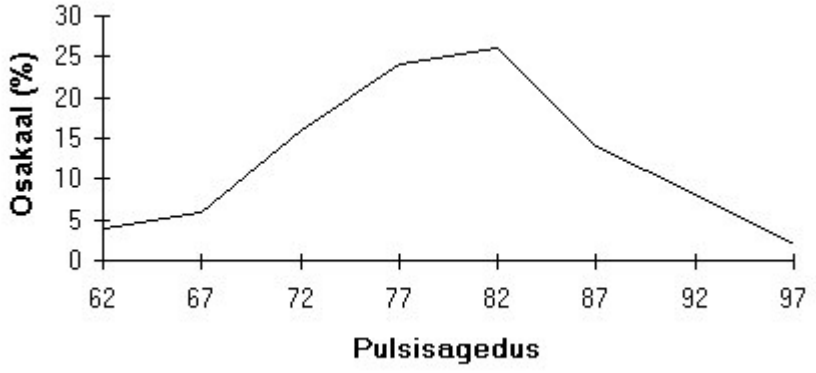
Arvutame nüüd välja iga vahemiku osakaalu protsentides ning esitame sagedustabeli nende osakaalude abil:

50 tudengi pulsisagedused (lööki minutis). SAGEDUSTABEL.

Pulsisagedus (lööki minutis)	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Vahemiku keskpunkt	62	67	72	77	82	87	92	97
Osakaal (%)	4	6	16	24	26	14	8	2

Sagedustabeli võib esitada ka teistsuguse joonise abil. Kanname iga vahemiku keskpunkti kohale punkti - punkti kõrguse määrab jällegi iga vahemiku osakaal - ning ühendame saadud punktid murdjoonega. Niisugust joonist nimetatakse JAOTUSPOLÜGOONIKS.

50 tudengi pulsisagedused (lööki minutis). JAOTUSOLÜGOON.



3.2 Keskmist tendentsi väljendavad arvkarakteristikud (keskmised).

Nagu te toodud näidete puhul olete märganud, on jaotuse 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 me nimetamegi keskmiseks tendentsiks. Eelmises peatükis tutvusime me juba kolme arvkarakteristikuga, 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	Seda liiki kasutavate õpilaste arv
Tramm	16
Buss	14
Jalgsi	9
Auto	6
Jalgratas	3
Mootorratas	2
Kokku	50

* * *

Sellise kategoriaalse tunnuse puhul saab keskmistest kasutada ainult moodi. Kõige populaarsem transpordivahend kooli jõudmiseks ehk mood on siin “tramm”.

Kui nominaalsete tunnuste puhul saab keskmist tendentsi väljendada ainult moodi abil, siis kvantitatiivsete andmete puhul on võimalik leida kõik kolm erinevat keskmist. Enam kasutatavateks on siiski keskväärtus (aritmeetiline keskmine) ja mediaan.

Kõige eelistatum keskmist tendentsi väljendav suurus on keskväärtus, sest ta on kõige stabiilsem st. võttes ühest üldkogumist erinevaid valimeid muutub keskväärtus mediaani ja moodiga võrreldes kõige vähem. Siit järeldus, et ta iseloomustab üldkogumit paremini kui mediaan või mood.

Sellegi poolest on situatsioone, kus keskmist tendentsi on õigem iseloomustada mediaani abil. Vaadake kahte alljärgnevat jaotust. Mõlemas on toodud viie inimese kuupalgad:

I	1000 kr.	1400 kr.	2000 kr.	2500 kr.	3000 kr.
II	1100 kr.	1400 kr.	1900 kr.	2600 kr.	13000 kr.

Mediaanid kahes grupis on küllalt sarnased: I - 2000 kr., II - 1900 kr. Arvutades aga välja keskväärtused saame, et keskväärtus esimeses grupis on 1980 krooni ning teises grupis 4000 krooni.

Esimest gruppi 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 kasutada mediaani, sest keskväärtus on tugevalt mõjutatud ühest ebatüüpilisest, teistest väga erinevast väärtusest, mediaani sellised ekstreemsed väärtused aga ei mõjuta.

Samuti tuleks kasutada mediaani, kui jaotuse mõned väärtused on täpselt teadmata. Näiteks olgu meil teada osakonna töötajate vanused järgmiselt:

alla 20,23, umbes 25, 30, 36, 42, üle 45 aasta.

Siin me ei saa arvutada keskväärtust, kuid mediaan on täpselt 30 st. pooled töötajad on nooremad ja pooled töötajad on vanemad kui 30 aastat.

Lõpetuseks toome ühe tabeli, mis näitab erinevate keskmiste kasutamise võimalikust eri tüüpi tunnuste puhul:

tunnuse tüüp keskmine	kategoriaalne nominaalne	järjestatud	kvantitatiivne diskreetne ja pidev
mood	JA	JA	JA
mediaan		JA	JA
aritmeetiline keskmine			JA

3.3 Hajuvust väljendavad arvkarakteristikud

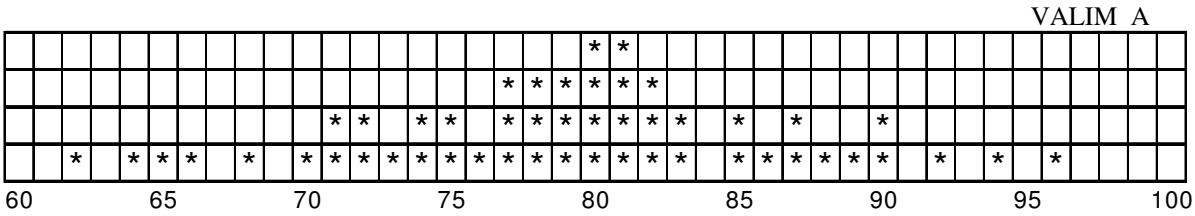
Alustame lihtsa näitega: olgu meil teada kahe üliõpilaste grupi kontrolltöö tulemused (hinnatud on 10-palli süstemis):

1) 6 7 7 7 8
2) 3 5 8 9 10

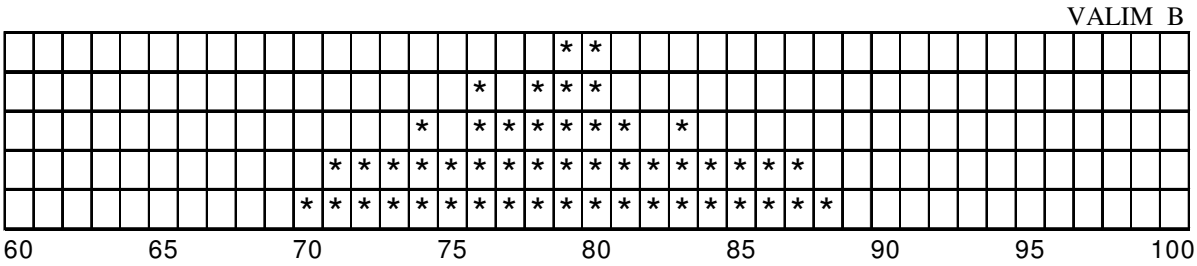
Mõlema grupi keskmiseks tulemuseks on 7 palli, kuid ometi näeme me selgelt kui suur on erinevus nende kahe grupi tulemuste vahel: esimese grupi üliõpilaste teadmiste tase on märksa ühtlasem kui teise grupi üliõpilastel st. esimese seeria väärtused koonduvad tihedalt keskmise ümber samal ajal kui teises seerias paiknevad väärtused hajusalt. Hajuvus ongi keskmise kõrval teine oluline variatsioonirida iseloomustav suurus.

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

50 tudengi pulsisagedused (lööki minutis)



50 tudengi pulsisagedused (lööki minutis)



Mis on teie arvates kõige suurem erinevus nende kahe jaotuse vahel? Kas te oskate öelda, milline juba õpitud arvkarakteristikutest aitab seda erinevust mõõta?

* * *

Diagrammidele peale vaadates võime kohe näha, et esimene jaotus on rohkem välja venitatud st. pulsisagedused valimis A on rohkem hajunud kui valimis B.

Jaotuse hajuvust ehk variatiivsust saame me kõige lihtsamini väljendada arvutades jaotuse ulatuse. Meie näites:

valimis A on ulatus = 96 - 62 = 34 lööki
valimis B on ulatus = 88 - 70 = 18 lööki.

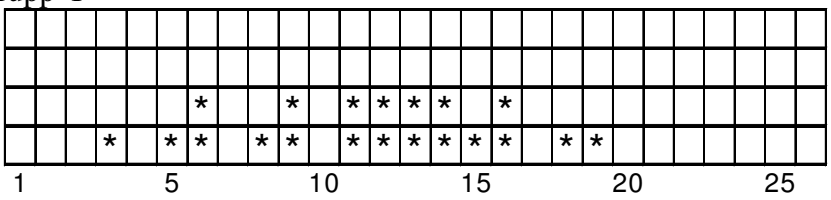
Seega valimis B on hajuvus palju väiksem kui valimis A.

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

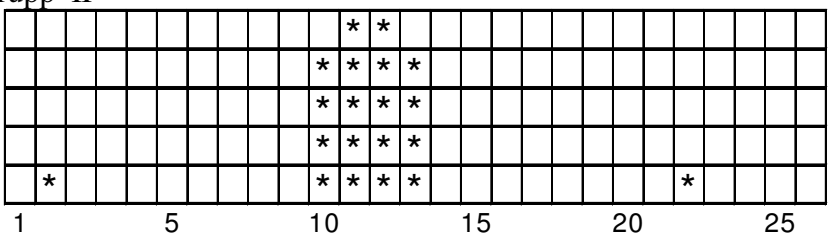
põhjusel olla teistest väga erinevad nn. ekstreemsed väärtused. Tuletame meelde näidet palkadest! Seepärast on selle näitaja usaldatavus väike ning teda kasutatakse vaid jaotusest kõige üldisema pildi saamiseks.

Vaatame veel kahte punktdiagrammi. Siin on kujutatud kahe üliõpilaste grupi (mõlemas 20 tudengit) testitulemused:

Grupp I



Grupp II



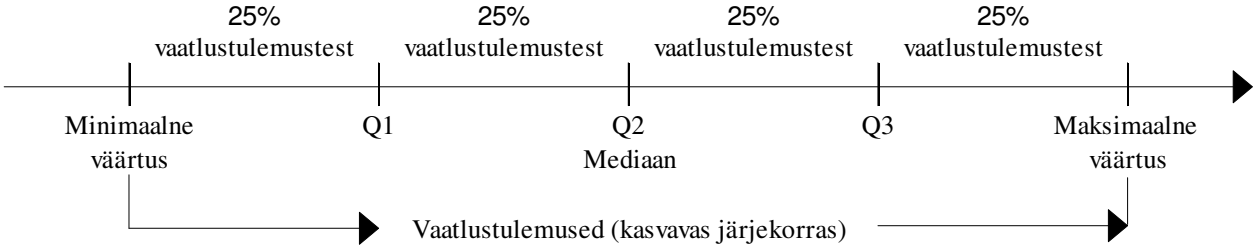
Kumb toodud jaotustest on teie arvates suurema hajuvusega? Kas ka ulatus selles jaotuses on suurem?

* * *

Üldiselt tundub, et esimese grupi hinnete hajuvus on suurem, sest selles grupis on saadud 13 erinevat hinnet, kusjuures teises grupis, kui välja jätta kaks ekstreemset väärtust, on hinded jaotunud väga ühtlaselt ainult nelja erineva hinde 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.

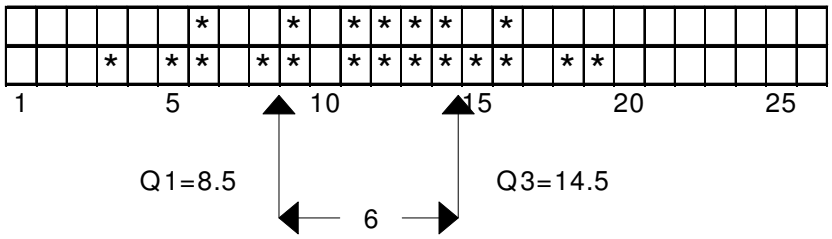
Sellise piirkonna moodustamisel on meile abiks KVARTIILID. Kui mediaan jagab meie vaatlustulemused kahte võrdsesse ossa, siis kvartiilid võimaldavad need jagada nelja võrdesse ossa nii, et igasse ossa jääb 25% tulemustest:



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 tudengit, 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?

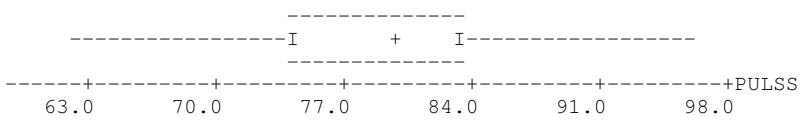
* * *

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.

Kvartiilide vahet kasutatakse tihti koos mediaaniga ning ta määrab ära vahemiku, milles asuvad pooled valimi elemendid.

Jaotuse hajuvust saab iseloomustada ka graafiliselt - karp-vurrud-diagrammi abil. Sellel joonisel esitatakse üheaegselt mitu erinevat arvkarakteristikut. Karp-vurrud-diagrammil kujutatakse kvartiilid (seega ka mediaan) 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.

50 tudengi pulsisagedused (KARP-VURRUD-DIAGRAMM)



Kõige sagedamini kasutatav hajuvuse näitaja on STANDARDHÄLVE. Nagu aritmeetiline keskmine, nii võtab ka standardhälve arvesse kõik vaatlustulemused. Kui meie vaatlustulemused on kõik ühesugused (N. kõik tudengid said kontrolltööl 12 palli), siis hajuvust ei ole ning aritmeetiline keskmine on võrdne selle sama väärtusega. Seega 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 aritmeetilisest keskmisest (keskmiselt) erinevad. Mida suurem on hajuvus, seda suuremad on erinevused ning seda suurem on ka standardhälve.

Kumba järgneva variatsioonirea puhul on teie arvates standardhälve suurem?

1)	6	24	37	49	64	(keskmine = 36)
2)	111	114	117	118	120	(keskmine = 116)

* * *

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

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

Väärtus:	111	114	117	118	120
Erinevus 116'st:	- 5	- 2	+1	+ 2	+ 4

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 raskusest tõstetakse kõik hälbed ruutu:

Hälve:	- 5	- 2	+ 1	+ 2	+ 4
Hälve ruudus:	25	4	1	4	16

Saadud ruuthälvete aritmeetilist keskmist nimetatakse DISPERSIOONIKS:

$$Dispersioon = \frac{25+4+1+4+16}{5} = \frac{50}{5} = 10$$

Dispersioon on arvkarakteristik, mida statistikas küllalt palju kasutatakse, kuid tal on igapäevase praktilise kasutamise jaoks üks tülikas puudus: kui vaatlustulemused (ja seega ka keskväärtus) olid näiteks ühikutes 'lööki minutis' või 'millimeetrit', siis dispersiooni ühikuks oleks 'lööki minutis 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:

$$Standardhälve\ (jaotus\ 2) = \sqrt{10} = 3.16$$

Viime läbi samad arvutused jaotuse 1) jaoks:

Väärtus:	6	24	37	49	64
Erinevus 36'st:	-30	-12	+1	+13	+28
Hälve ruudus:	900	144	1	169	784

$$Dispersioon = \frac{900+144+1+169+784}{5} = \frac{1998}{5} = 399.6$$

$$Standardhälve\ (1) = \sqrt{399.6} = 20$$

Nagu te arvata võisite, on esimese jaotuse standardhälve palju suurem kui teise. See on nii sellepärast, et esimene jaotus on palju rohkem hajunud.

Vaatame nüüd uuesti selle peatüki alguses tutvustatud kahte punktdiagrammi. Üks alljärgnevatest arvupaaridest näitab neile jaotustele vastavaid standardhälbeid. Milline on teie arvates õige paar? Milline standardhälve kuulub millisele jaotusele?

a) 4.6 ja 7.6 lööki minutis,
b) 7.6 ja 37 lööki minutis või
c) 19 ja 37 lööki minutis.

* * *

Valimi A standardhälve on 7.6 ja valimi B 4.6 lööki minutis. Ma arvan, et teil ei tekkinud raskusi otsustamisel, et suurem standardhälve kuulub valimile A ja väiksem valimile B. Samuti ei tohiks olla raske näha, et paarides b) ja c) pakutud väärtus(ed) on suuremad kui jaotuste ulatused ning ei sobi seega standardhälbeks.

Tavaliselt ei ületa standardhälve poolt jaotuse ulatusest. Kui valimis on 10 objekti, siis võib oodata standardhälvet, mis on lähedane ühele kolmandikule jaotuse ulatusest. Kuid 100 liikmelise valimi puhul on oodatav standardhälve veel väiksem: umbes üks viiendik jaotuse ulatusest.

3.4 Kokkuvõte

Siiani oleme me rääkinud sellest, kuidas kogutud vaatlustulemustest ülevaadet saada ning kuidas nende andmete põhjal valimit kirjeldada. Tuletame nüüd meelde, mida me selleks saame teha:

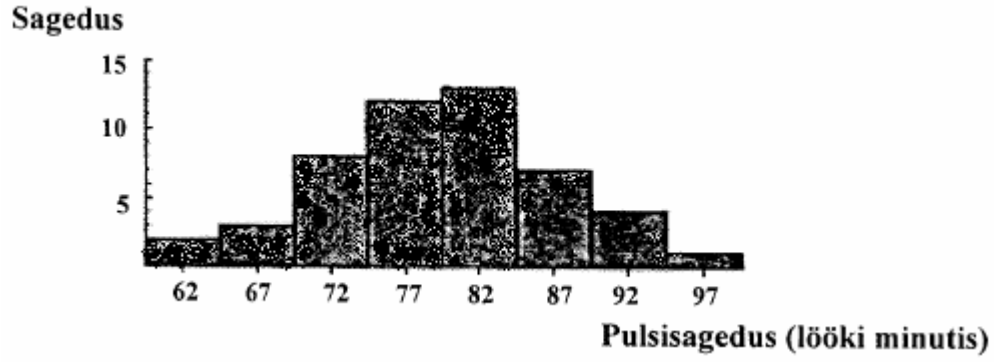
1. Paigutada andmed mõnda tabelisse, mis näitab andmete jaotumist kategooriate või minimaalse ja maksimaalse väärtuse vahel.
   (andmetabel, sagedustabel, variatsioonirida)
2. Illustreerida seda jaotumist diagrammi abil.
   (tulpdiagramm, sektordiagramm, punktdiagramm, histogramm, sageduspolügoon jne.)
3. Leida sobiv(ad) arvkarakteristik(ud) iseloomustamaks jaotuse ‘keskmist tendentsi’.
   (mood, mediaan, aritmeetiline keskmine)
4. Kvantitatiivsete andmete puhul leida ka hajuvust iseloomustav(ad) arvkarakteristik(ud) ning illustreerida hajuvust karp-vurrud-diagrammi abil.
   (ulatus, kvartiilide vahe, standardhälve).

Millist tabelit, diagrammi või arvkarakteristikut kasutada sõltub peamiselt sellest, kas andmed on kategoriaalsed või kvantitatiivsed.

4. Jaotuse kuju.

Aritmeetiline keskmine ja standardhälve võimaldavad meil ilmekalt kirjeldada kvantitatiivsetest andmetest koosnevaid statistilisi jaotusi; mediaan ja ulatus/kvartiilide vahe on sagedamini kasutatavad järjestustunnuse puhul või kui jaotuses leidub ekstreemseid, teistest oluliselt erinevaid väärtusi. Kuid nagu me nägime on jaotuse kirjeldamisel oluline osa ka joonistel ja graafikutel, mis võimaldavad luua parema ettekujutuse jaotuse üldisest *kujust*.

Kas olete märganud, et paljude tunnuste puhul on jaotus enam-vähem *sümmeetriline*? See tähendab, et kõige rohkem mõõtmistulemusi asub ulatuse keskosas ning liikudes ulatuse otspunktide poole mõõtmistulemuste hulk aina väheneb. Siin on üks näide sellisest jaotusest:



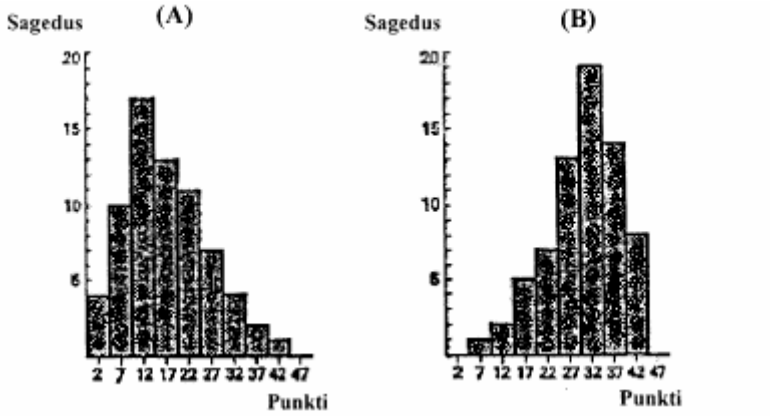
Selline sümmeetrilisus on statistiliste jaotuste puhul väga tavaline - eriti kui on tegemist bioloogiliste (ka psühholoogiliste) nähtustega. Kuid ta ei ole universaalne, ainuvõimalik.

4.1 Asümmeetrilised jaotused.

Vaadake alljärgnevat tabelit, milles on toodud samade õpilaste poolt kahel erineval matemaatika testil saadud punktid.

Punktid	Õpilaste arv	
	Test X	Test Y
0-4	4	0
5-9	10	1
10-14	17	2
15-19	13	5
20-24	11	7
25-29	7	13
30-34	4	19
35-39	2	14
40-44	1	8

Kumb järgnevatest histogrammidest illustreerib jaotust X ning kumb jaotust Y? Milles seisneb nende kahe jaotuse erinevus?



* * *

Vasakpoolne histogramm (A) vastab testi X tulemuste jaotusele ning parempoolne histogramm (B) vastab testi Y tulemuste jaotusele.

Selgelt on näha, et need jaotused ei ole sümmeetrilised: testitulemuste enamus (ning ka modaalne klass) ei asu ulatuse keskosas (st. kuskil 20 ja 30 punkti vahel) ning väärtused ei jagune kaugeltki mitte võrdselt kummalegi poole ulatuse keskpunkti. On aga näha, et need jaotused on välja venitatud erinevates suundades. Sellist jaotuse väljavenitatust nimetatakse *asümmeetriaks*. Kui jaotus on välja venitatud paremalt st. jaotuse "saba" jääb paremale (positiivsele) poole siis on tegemist positiivse asümmeetriaga, kui "saba" jääb aga vasakule poole, siis on tegemist negatiivse asümmeetriaga.

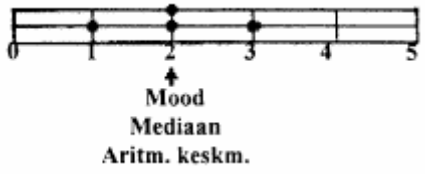
Kumba ülaltoodud jaotuse asümmeetria on positiivne ja kumba negatiivne?

* * *

Jaotuse (A) "saba" on paremal pool ning seega on tema asümmeetria positiivne, seevastu jaotuse (B) "saba" on vasakul pool ning asümmeetria negatiivne.

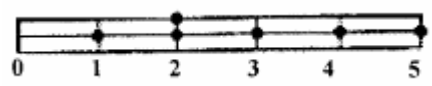
Kuigi jaotuse minimaalne ja maksimaalne väärtus on kumbagi jaotuse puhul küllalt sarnased, erinevad nende jaotuste keskväärtused üksteisest tunduvalt. Testi X puhul on tulemuste keskväärtus 17,1 punkti; testi Y puhul aga 30,2 punkti.

Vaatame järgnevalt millist mõju avaldab asümmeetria erinevatele keskmist tendentsi väljendavatel arvkarakteristikutele, nende suhtelisele suurusele ning omavahelisele paiknemisele. Sümmeetriliste jaotuste puhul asuvad kõik kolm keskmist: mood, mediaan ning keskväärtus ühes kohas - ulatuse keskosas. Vaadake näiteks järgmist lihtsat punktdiagrammi:

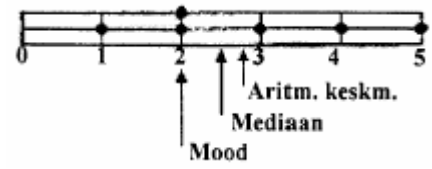


Sellel on üks tulemus väärtusega 1, kaks tulemust väärtusega 2 ning üks tulemus väärtusega 3. Mood ehk kõige sagedamini esinev väärtus on siin 2. Mediaan, mis jagab jaotuse kaheks võrdseks osaks on 2 ning aritmeetiline keskmine (1+2+2+3)/4 on samuti täpselt 2.

Vaatame nüüd, mis juhtub, kui me lisame sellele jaotusele "saba" st. lisame ühele poole mõned tulemused (näiteks tulemused väärtustega neli ja viis).

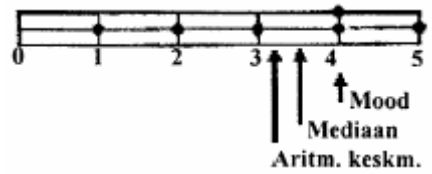


Kuidas see mõjutab kolme keskmist? Mood on ikka 2, kuid mis juhtus mediaaniga? Kuna nüüd on meil kuus tulemust siis mediaan peab asuma kolmanda ja neljanda väärtuse vahel. Kolmas väärtus on 2 ja neljas 3, seega mediaan on nüüd 2,5. Näeme, et mediaan nihkus moodist eemale "saba" suunas.
Mis juhtus aga keskväärtusega?
* * *
Arvutades välja keskväärtuse (2,8) näeme, et see on nihkunud samuti "saba" suunas ning isegi rohkem kui mediaan. Seega paiknevad asümmeetrilises jaotuses keskväärtused järgmiselt:

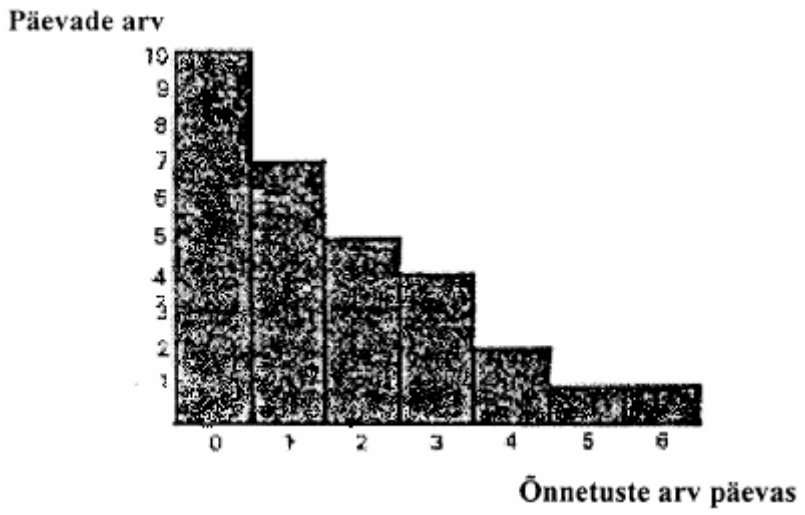


Juhul kui jaotuse "saba" oleks olnud teisele poole, kuidas oleks nihkunud mediaan ja keskväärtus võrreldes moodiga siis?
* * *

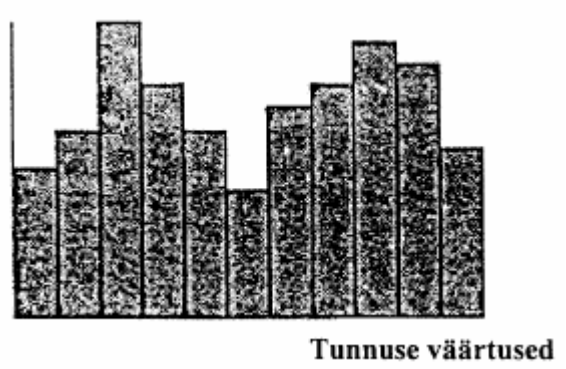
Vastus peitub järgmisel joonisel:



Asümmeetrilise jaotuse puhul võib erinevate keskmiste omavahelise paiknemise alati ette ennustada. Keskväärtus on moodist (jaotuse tipust) nihutatud alati "saba" suunas ning mediaan asub nende kahe vahel. Mida suurem asümmeetria, seda suurem vahemaa jääb moodi ja keskväärtuse vahele. (Jaotuse asümmeetria iseloomustamiseks kasutatakse arvkarakteristikut, mida nimetatakse *asümmeetria koefitsiendiks - coefficient of skewness*.)
Reaalselt kogutud andmete puhul on jaotus tavaliselt alati mingil määral asümmeetriline. Kuid leidub jaotusi, mille puhul asümmeetria on veelgi suurem kui eelpool toodud näidetes. Järgmisel joonisel on illustreeritud, vaatlustulemusi, mis on saadud ühe kindla ristmiku jälgimisel. Nagu te näete on selliseid päevi, mil pole juhtunud ühtki liiklusõnnetust, enam kui selliseid, mil on toimunud üks õnnetus jne.



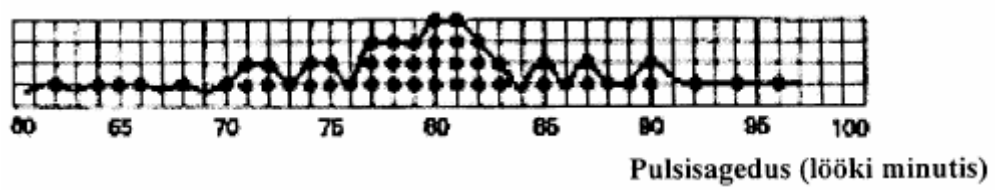
Tegelikust elust võib tuua veel teisigi näiteid suure asümmeetriaga jaotustest. Tüüpiliselt saame me sellise jaotuse palkade ning suremuse puhul, kuid ka näiteks mingi uue toote katsepartii testimisel, kus arvestatakse tootes esinenud vigade arvu jne.
Nagu te edaspidi näete põhinevad paljud järeldava statistika meetodid eeldusel, et tegemist on enam-vähem sümmeetrilise jaotusega, seepärast on oluline enne analüüsimeetodi valimist pöörata tähelepanu ka jaotuse kujule. Kuid ka sümmeetrilise jaotuse puhul võib esineda tavalisest (pean siin silmas jaotust, kus mõõtmistulemused koonduvad ümber ulatuse keskpunkti) erinevaid jaotuse kujusid. Näiteks võib meil olla tegemist *bimodaalse* sümmeetriaga, kus jaotusel on kaks tippu, mis asetsevad üks ühel ja teine teisel pool ulatuse keskpunkti (vt. joonist). Sellisel juhul on tavaliselt tegemist olukorraga, kus on üheaegselt mõõdetud kahte väga erinevat gruppi ning seetõttu väljendab joonis tegelikult kahte osaliselt kattuvat jaotust.



Kas oskad tuua näiteid olukordadest, kus võib saada bimodaalse jaotuse?
* * *

4.2 Normaaljaotuse idee

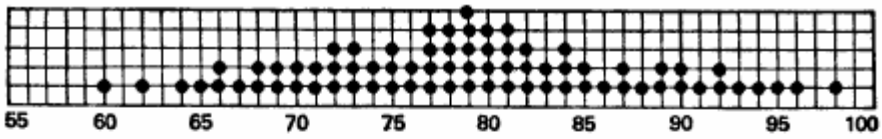
Enne kui me jätkame jaotuse kuju uurimist, vaatame kuidas saaks jaotust lihtsamalt, ilma punktide ja tulpadeta, kujutada. Selleks ühendame ülemised punktid punktdiagrammil või tulpade keskpunktid tulpdiagrammil sujuva kõverjoonega (mitte sirglõikudega nagu me tegime jaotuspolügooni saamiseks). Nii saame kõverjoone, mida nimetatakse JAOTUSKÕVERAKS. Alljärgnevalt jooniselt on näha kuidas käib jaotuskõvera konstrueerimine:



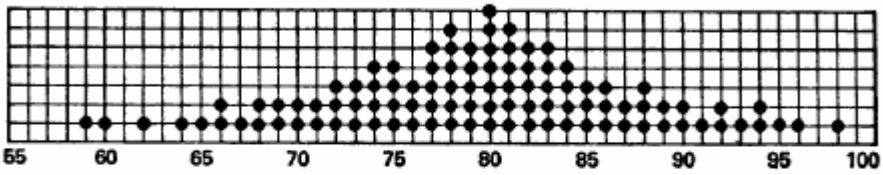
Viimasel joonisel on kujutatud jällegi teile juba tuttav jaotus 50 tudengi pulsisagedustest. Kuigi jooniselt on näha, et pulsisagedustel on kalduvus koonduda ulatuse keskossa, ei ole see koondumine sugugi ühtlane ning jaotuskõver on küllalt sik-sakiline.
Mis te arvate, mis juhtub jaotuskõveraga, kui sinna lisada veel näiteks viiekümne tudengi pulsisagedused?
* * *

Kui te arvasite, et jaotuskõver muutub siledamaks, siis oli teil õigus. Tõepoolest, mida enam vaatlustulemusi me joonisele kanname (st. mida suurem on meie valim), seda siledam on jaotuskõver. Kas oskad seda nähtust oma sõnadega põhjendada?
Kirjeldatud efekti iseloomustab järgmine jooniste seeria:

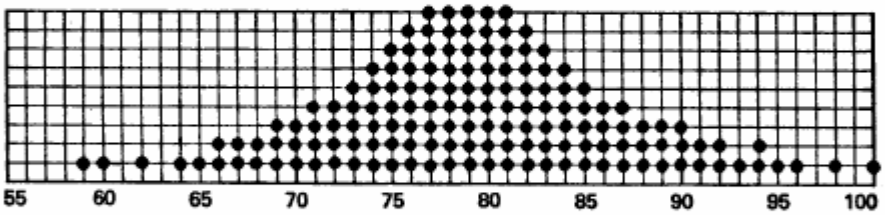
75 tudengit



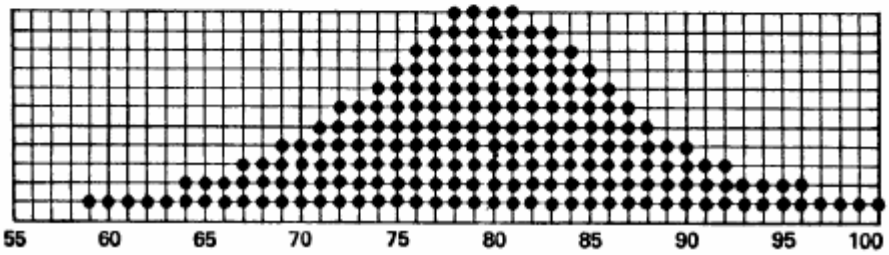
100 tudengit



150 tudengit

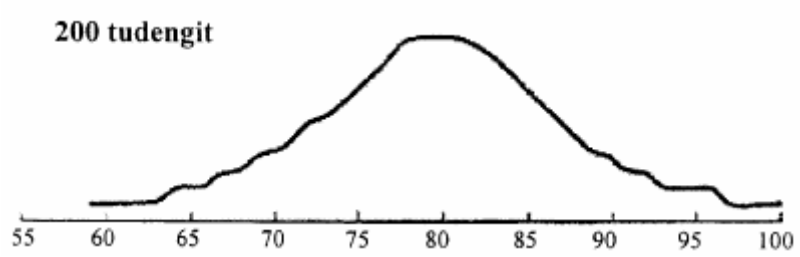


200 tudengit

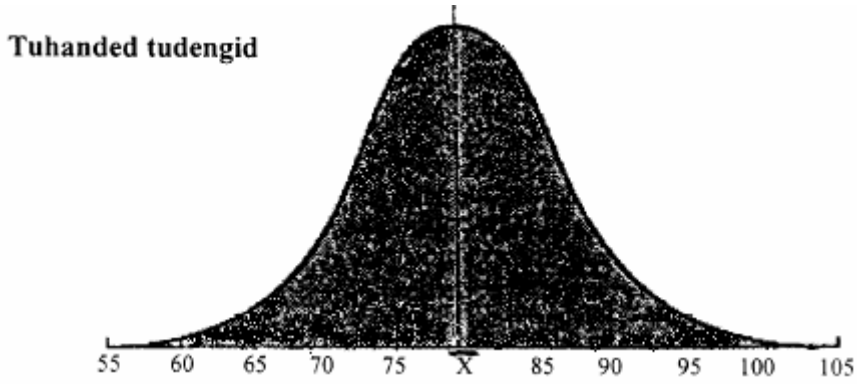


Joonistage välja jaotuskõver viimase joonise jaoks, kus on toodud 200 tudengi pulsisagedused.
* * *

Saadud kõver peaks välja nägema umbes nii:



Jaotuskõver on nüüd märksa siledam kui esialgsel joonisel, kus oli toodud vaid 50 tudengi pulsisagedused. Kui aga kujutada ette olukorda, kus oleks võimalik mõõta mitte sadade vaid *tuhandete* tudengite pulsisagedus, siis peaks jaotuskõver hakkama lähenema sellisele ideaalselt siledale kõverale:



Ülaltoodud jaotuse kuju sarnaneb kõveraga, mida nimetatakse NORMAALJAOTUS-KÕVERAKS. Normaaljaotuskõver on kellukakujuline, ideaalselt sümmeetriline ulatuse keskpunkti suhtes ning nii keskväärtus, mediaan kui ka mood asuvad kõik täpselt ulatuse keskpunktis. (Esimesena kasutas sellise kõvera ideed Inglise matemaatik de Moivre kuueteistkümnendal sajandil.)
Pange tähele, et mistahes jaotuskõvera alla peab alati mahtuma 100% vaatlustulemustest ning mistahes vertikaaljoon, mis on tõmmatud horisontaalteljelt üles kuni kõverani jaotab kõvera poolt moodustatud piirkonna kaheks osaks. Nagu te õige pea näete, saab normaaljaotuse puhul leida iga konkreetse väärtuse jaoks, kui mitu protsenti vaatlustulemustest jääb temast vasakule ja kui mitu paremale poole.
Vaadake näiteks ülaltoodud joonist; kas te oskate öelda mitu protsenti vaatlustulemustest jääb kummalegi poole keskväärtust?
* * *

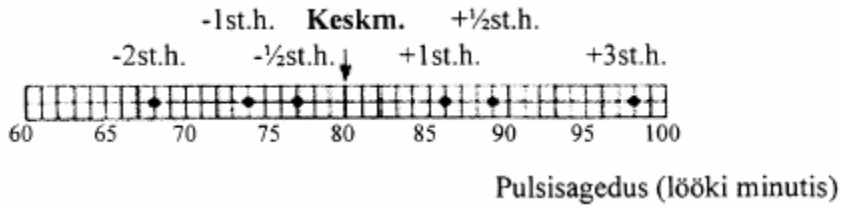
Kuna tegemist oli ideaalselt sümmeetrilise jaotusega, siis jääb kummalegi poole 50% tulemustest.

Rääkides normaaljaotuskõverast, ei kasutata sõna ‘normaalne’ tähenduses ‘tavaline’ vaid pigem tähistab see ‘standardset’ ‘ideaalset’ kõverat, millega meil on võimalik tegelikke jaotusi võrrelda. Normaaljaotuskõvera idee kasulikkus seisneb selles, et väga paljud tegelikus elus ette tulevat

tunnused jaotuvad piisavalt suure vaatluste arvu juures ligilähedaselt normaalsele: st. nad on küllalt kellukakujulised ning sümmeetrilised keskpunkti suhtes. Seega on normaaljaotuskõver matemaatiline abstraktsioon, mis on defineeritud väga keerulise võrrandiga ning eeldab üldkogumit, mis oleks lõpmatult suur.
Vaatamata sellele võime tegelikus elus ka väikeste valimite puhul saada küllalt kellukakujulisi jaotusi. See annab põhjust oletada, et meie väike valim on pärit suurest üldkogumist, mida võib kirjeldada normaaljaotuskõveraga.
Kui see on nii, siis saame valimi mõõtmistulemuste interpreteerimisel kasutada mõningaid väga kasulikke normaaljaotuse omadusi. Normaaljaotuse kuju on selline, et *me saame alati leida kui mitu protsenti üldkogumist jääb suvalise kahe väärtuse vahelisse lõiku.* Selleks on meil vaja teada vaid jaotuse keskväärtust ning standardhälvet. Teades neid kahte näitajat, saame jaotuse iga üksiku väärtuse puhul öelda, et ta asub keskväärtusest 'nii mitme standardhälbe' kaugusel. Seega me võime standardhälvet kasutada kui *mõõtühikut*.

Oletame näiteks, et meil on tudengite pulsisageduste valim, mille keskväärtus on 80 ja standardhälve 6 lööki minutis. Tudeng, kelle pulsisagedus on 86 lööki minutis asub siis 'ühe standardhälbe võrra ülalpool keskmist'. Samuti tudeng, kelle pulsisagedus on 71 lööki minutis asub keskmisest 1,5 standardhälbe võrra madalamal jne.
Kus asuksid siis pulsisagedused 74, 89, 98, 77 ja 68 lööki minutis?
* * *

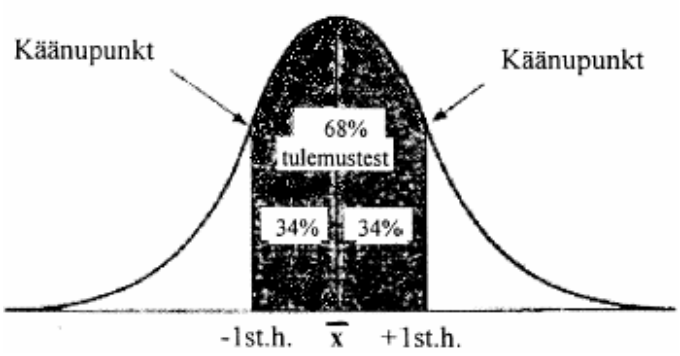
Vaatame nüüd kuidas seda kujutada pulsisageduste skaalal:



Seega *igat* jaotuse väärtust on võimalik väljendada kui nii mitme standardhälbe võrra all- või ülalpool keskväärtust asuvat. Pange tähele, et seda saame me teha vaatamata sellele, kas jaotus on normaalne või mitte. Kui meil on aga tegemist normaalse (või peaaegu normaalse) jaotusega, siis võime normaaljaotuse omadusi kasutades leida mistahes kahe väärtuse jaoks kui suur osa mõõtmistulemustest jääb nende vahele.

4.3 Proportsioonid normaaljaotuskõvera all.

Vaadake allpool toodud normaaljaotuskõverat. Langedes tipust kumalegi poole on kõver alguses kumer minnes ühes kindlas punktis üle nõgusaks. Seda punkti nimetatakse käänupunktiks ning ta asub täpselt ühe standardhälbe kaugusel keskväärtusest. Umbes 2/3 (68%) kõigist vaatlustulemustest jäävad vahemikku, mis hõlmab ühe standardhälbe mõlemalt poolt keskmist. Tähistame seda vahemikku $x \pm 1$st.h.



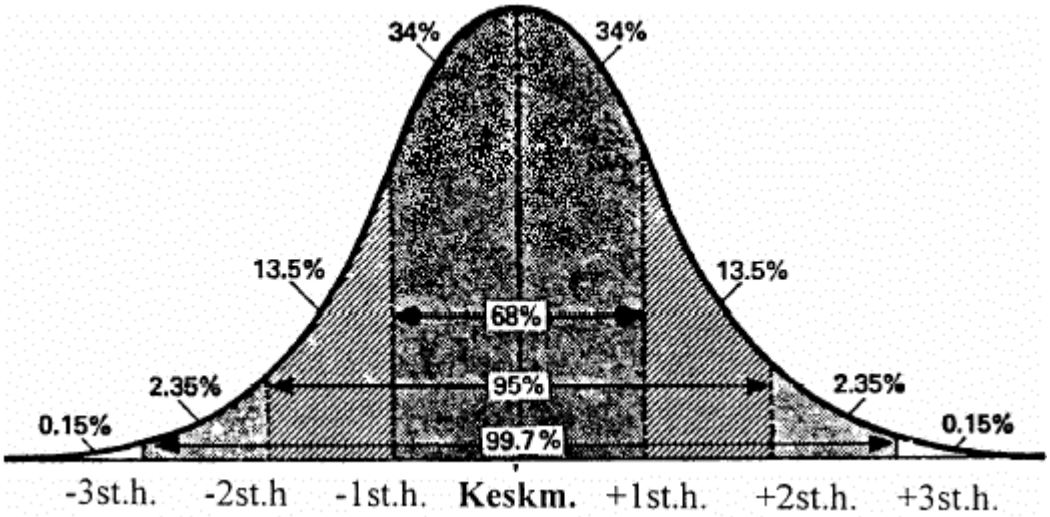
Kui suur osa vaatlustulemustest on normaaljaotuse puhul suuremad kui 1 st.h.?
* * *

Väljaspool vahemikku x ± 1st.h. asub 100-68=32% tulemustest, kuna jaotus on sümmeetriline, siis saame, et 16% tulemustest on suuremad kui x ± 1st.h.

Mis meil sellest teadmisest reaalses elus kasu võiks olla? Toon ühe lihtsa näite. Kloostrimetsa teel mõõtsid politseinikud möödasõitvate autode kiirust (kokku umbes 1000 autot). Osutus, et kiiruste jaotus oli ligilähedane normaaljaotusele. Keskmiseks kiiruseks saadi 55 km tunnis ning kiiruste standardhälve oli 5 km/h. Nagu te teate on Tallinnas lubatud sõidukiirus 50 km/h; kas te oskate ligikaudselt öelda, kui suur osa möödasõitnud autodest rikkus seadust?
* * *

Tuleb välja, et ligikaudu 84 % sõidukitest on ületanud enamal või vähemal määral lubatud kiirust.

Nii siis nägime, et vahemikku x ± 1st.h. jääb ligikaudu 68% tulemustest. Kui palju jääb aga vahemikku x ± 2st.h. On võimalik välja arvutada, et see piirkond sisaldab endas 95% tulemustest. Vahemik x ± 3st.h. hõlmab aga peaaegu kõik vaatlustulemused ehk 99,7%.
Nagu juba öeldud, on normaaljaotuse puhul tegelikult võimalik välja arvutada mistahes vahemikku jäävate tulemuste osakaal. Ainuke operatsioon, mis meil tuleb teha on 'tõlkida' oma esialgsed tulemused (näiteks 58 km/tunnis, 85 lööki minutis või 96 punkti) ühikutesse 'nii mitu standardhälvet all- või ülalpool keskväärtust'.



Oletame näiteks, et Eesti koolides on läbi viidud ajaloo ainetest. Tulemused on saadud 650 õpilase testimisel ning tulemuste jaotus vastab normaaljaotusele. Keskmine testitulemus on 78 punkti ning standardhälve 6 punkti. Juku sai testil 81 punkti. On ilmne, et tema tulemus on üle keskmise, kuid kuidas saaks seda väljendada üldisemalt - võttes mõõtühikuks standardhälbe?

* * *

Tõepoolest, arvutus on lihtne: Juku testitulemusest tuleb lahutada keskmine tulemus ning saadud erinevus jagada ühiku ehk standardhälbega. Seega, Juku tulemus on keskmisest kõrgem (81-78)/6=3/6=0,5 standardhälbe võrra. Vaatlustulemustele vastavaid väärtusi, kus ühikuks on standardhälve nimetatakse tihti *z-väärtusteks*.

Leia sama jaotuse korral tulemustele 72, 90, 63 ja 96 punkti vastavad z-väärtused.
* * *

4.4 Erinevate tunnuste väärtuste võrdlemine

5. Valimilt üldkogumile ehk järelduste tegemine üldkogumi kohta valimi põhjal

5.1 Valimi moodustamine

Kursuse esimeses pooles vaatasime kuidas kirjeldada meie poolt kogutud andmeid st. tutvusime erinevate meetoditega, mis võimaldavad iseloomustada neid objekte (katseisikuid) mida me mõõtnud või vaadelnud oleme. Tihti on aga uuringute eesmärgiks mingi laiema objektide hulga ehk ÜLDKOGUMI kirjeldamine, mille kõiki objekte ei ole reaalselt võimalik (ega ka mõttekas) vaadelda. Seega tuleb mõõtmiseks välja valida mõned üldkogumi objektid, mis esindaksid üldkogumit st. peegeldaksid piisavalt hästi üldkogumile iseloomulikke jooni. Need mõõtmiseks välja valitud objektid moodustavad VALIMI.
Selleks, et valim annaks üldkogumi kohta objektiivset ja usaldatavat informatsiooni, tuleb valimi liikmed valida JUHUSLIKULT. 'Juhuslikult' ei ole sugugi mitte sünonüüm 'suvalisele'; juhuslikkus statistikas tähendab, et igal üldkogumi liikmel peab olema võrdne võimalus valimisse valitud saada. Selleks, et tagada valimi juhuslikkus on mitmeid võimalusi. Vaatame nendest mõnda lihtsamat.

Kui meil on kasutada üldkogumi liikmete nimekiri, siis tuleb kõigepealt üksikud liikmed nummerdada ning määrata valimi maht st. valimisse valitavate objektide arv. Kui see on tehtud, siis tuleb juhuslike arvude tabeli või arvuti juhuslike arvude generaatori abil leida vastav hulk arve, mis määravad ära valimisse valitavate üldkogumi liikmete järjekorranumbrid. Sellist valimi moodustamise meetodi nimetatakse JUHUVÄLJAVÕTUKS.
Sageli pole aga meie käsutuses sellist üldkogumi liikmete nimekirja. Oletame näiteks, et te viite läbi tänavaküsitlust. Te ei tea, kes teile järgmisena vastu tuleb ning loomulikult ei saa te neid vastutulijaid eelnevalt nummerdada. Kas te näete mingit võimalust, kuidas igal tänaval liikujal oleks võrdne võimalus teie valimisse sattuda?
* * *

Teil tuleks eelnevalt otsustada, et te küsitlete peale iga eelmise intervjuu lõppemist näiteks täpselt viiendat vastutulijat või siis inimest, kes tuleb teile vastu täpselt 1 (või 2, või 3 või jne.) minutit peale eelmise intervjuu lõppu. Sellist meetodi nimetatakse SÜSTEMAATILISEKS VÄLJAVÕTUKS. Süstemaatilist väljavõttu kasutatakse vahel ka siis kui üldkogumi liikmete nimekiri on olemas. Sel juhul valitakse nimekirjast välja vastavalt kas iga viies, kümnes, kolmeteistkümnes või jne. objekt.

Mõnel juhul, kui on ette teada, et üldkogum koosneb erinevatest osadest (näiteks kõrgkoolis õpib 700 naissoost ja 300 meessoost tudengit) ning meil on põhjust arvata, et need osad omavahel mõne tunnuse osas erinevad, siis on mõistlik kasutada TÜÜP- e. SEERIAVÄLJAVÕTTU, kus eelnevalt otsustatakse kui palju liikmeid valitakse valimisse igast üldkogumi erinevast osast. Tavaliselt tehakse seda proportsionaalselt üldkogumi tegeliku jaotusega. Seega, kui me tahame saada sajaliikmelist valimit, mis oleks proportsionaalne eelnevas näites toodud üldkogumiga, siis peaksime välja valima 70 naist ja 30 meest, kusjuures naiste ja meeste hulgast tuleb valimi liikmed valida eelpool toodud nõudeid arvestades st. juhuslikult.

5.2 Järeldamine statistikas.

Niisiis teame me, et üldkogumit saab objektiivselt iseloomustada vaid juhuslik valim. Tuletame nüüd aga uuesti meelde näite, kus oli mõõdetud 50 tudengi pulsisagedused. Olgu meile teada, et need 50 katseisikuks olnud tudengit valiti kõigi tudengite seast välja juhuslikult. Eelpool arvutasime välja nende tudengite keskmise pulsisageduse, milleks on 79,1 lööki/minutis. Kas võime teha järelduse, et ka kõigi tudengite (see tähendab kogu üldkogumi) mõõtmisel saadaks keskmiseks pulsisageduseks täpselt 79,1 lööki/minutis?
* * *

Loomulikult mitte! Kuid me võime väita, et need kaks keskmist on küllalt sarnased. Edasi peaksimegi küsima: kui sarnased, kui lähedased nad siis üksteisele on?
Valimi põhjal arvutatud arvkarakteristikud (näiteks aritmeetiline keskmine, ulatus standardhälve jne.) on HINNANGUTEKS vastavatele üldkogumi parameetritele. Kui valimi arvkarakteristikuid tähistatakse tavaliselt Rooma tähtedega (näiteks valimi keskväärtust $\bar{x}$), siis üldkogumi parameetreid tähistatakse Kreeka tähtedega (näiteks üldkogumi keskväärtust μ).
Seega teades valimi keskväärtust saame me hinnata üldkogumi keskväärtust; valimi ulatus võimaldab leida üldkogumi ulatuse, valimi standardhälve annab meile ettekujutuse üldkogumi standardhälbest jne. Sellist hinnangute andmist üldkogumi parameetrite kohta nimetataksegi STATISTILISEKS JÄRELDAMISEKS.
Mis te arvate, millest sõltub statistilise järeldamise täpsus?
* * *

Mida rohkem on meie käsutuses informatsiooni, seda täpsemad on meie järeldused ehk mida suurem on valim, seda täpsem on meie hinnang üldkogumile. Edaspidi näeme, et peale valimi suuruse on veel teisigi tegureid, mis mõjutavad hinnangute täpsust, kuid põhiline ja ühtlasi meie poolt mõjutatav on just valimi suurus.
Seega, suurendades valimit saame suurendada oma järelduse täpsust, kuid me ei saa kunagi öelda, et üldkogumi keskväärtus (või mediaan või standardhälve) on 100%-lise kindlusega võrdne ühe konkreetse arvnäitajaga. (Välja arvatud juhul, kui me mõõdame kõiki üldkogumi elemente. Seda olukorda käsitlesime juba eespool ning seetõttu jätame selle võimaluse praegu kõrvale.) Parim, mis me teha saame, on väita, et "õige" üldkogumi keskväärtus või üldkogumi standardhälve või mistahes teine üldkogumi arvkarakteristik asub ühe või teise tõenäosusega ühes või teises väärtuste vahemikus.
Statistiline järeldamine on alati seotud veaga, mida ükski valem ega statistiline meetod ei suuda kõrvaldada. Küll aga võimaldavad viimased meil seda viga hinnata - mõõta.

5.3 Valimite keskväärtuste jaotus

Vaatame ühte näitlikku olukorda. Oletame, et teie rühma tudengitele tehti ülesandeks uurida kuivõrd rahul on erialakaaslased õpingutingimustega meie ülikoolis. Lihtsuse mõttes piirdume ühe küsimusega:

> Hinnake oma rahulolu õpingutingimustega meie ülikoolis kümnepalli skaalal, kusjuures 1 = pole absoluutselt rahul ning 10 = olen täiesti rahul.

Iga tudeng pidi uurimuse läbi viima iseseisvalt. Selleks pidi igaüks koostama kõigist vastaval erialal õppivatest tudengitest kahekümneliikmelise juhusliku valimi ning seejärel küsitluse läbi viima. Kui nüüd igaüks arvutaks välja saadud vastuste keskväärtuse, kas need oleks kõik ühesugused?

* * *

Ei! Teie erinevate valimite keskväärtused oleks mingil määral üksteisest erinevad. Tuletage meelde, et üldkogum oli kõigil sama: kõik teie erialal õppivad tudengid.
Mida suuremad valimid te moodustaksite, seda vähem erineksid teie valimite keskväärtused teineteisest, kuid erinevus ei muutuks olematuks enne kui igaüks teist küsiks arvamust iga erialakaaslase käest (see tähendab küsitleks kogu üldkogumit).

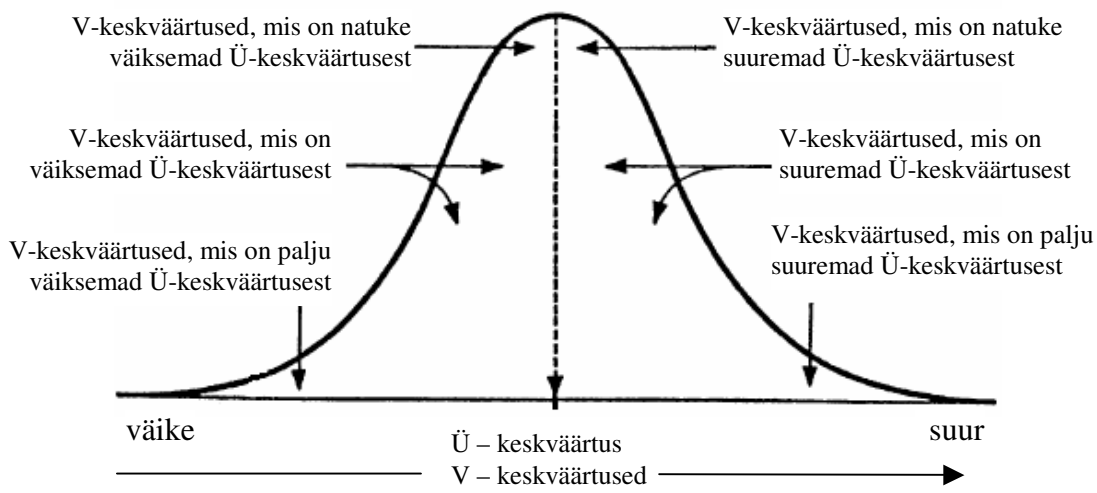
Nagu te näete erinevad ühe üldkogumi põhjal moodustatud valimite keskväärtused teineteisest. Me võime neid valimite keskväärtusi vaadata kui uut statistilist *tunnust*:

> valimi A keskväärtus (näiteks 6,7 palli)
> valimi B keskväärtus (7,3 palli)
> . . .
> valimi N keskväärtus (7,1 palli)

Kuigi tegelikus elus (tegeliku uurimuse puhul) moodustame me üldkogumi iseloomustamiseks ühe valimi, on järeldava statistika põhiolemuse mõistmiseks vaja ette kujutada just sellist olukorda, kus meil on võimalik samast üldkogumist võtta suur hulk etteantud suurusega juhuslikke valimeid. Iga erineva valimi jaoks saame leida tema keskväärtuse. Edasi vaatamegi neid **valimite keskväärtusi**. Me saame nendest keskväärtustest moodustada sagedusjaotuse st. me saame neid keskväärtusi järjestada ning võrrelda nende suuruse järgi. Seega oleme saanud VALIMITE KESKVÄÄRTUSTE JAOTUSE.
Valimite keskväärtuste jaotusel on oma keskväärtus (mis saadakse arvutades aritmeetiline keskmine üksikute valimite keskväärtustest) ning standardhälve. Kui meil on võetud piisavalt suur arv valimeid, siis koonduvad valimite keskväärtused ümber *üldkogumi keskväärtuse*. Seega valimite keskväärtuste keskväärtus ongi võrdne üldkogumi keskväärtusega.
Mis te arvate, milline kuju on valimite keskväärtuste jaotusel?

* * *

Nagu te võisite arvata, vastab valimite keskväärtuste jaotus normaaljaotusele. Isegi siis kui üldkogum ise (ja seega ka valimid) ei vasta normaaljaotusele, vastab valimite keskväärtuste jaotus ikkagi enam-vähem normaaljaotusele. Mida suuremad on valimid, seda sümmeetrilisem ja kellukakujulisem on nende keskväärtuste jaotus.

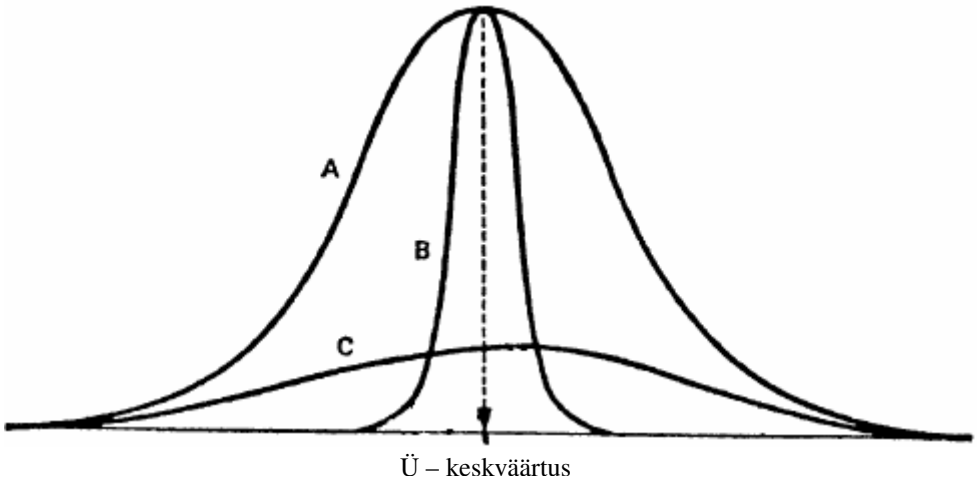


Tegelikus elus ei kohta me sellist valimite keskväärtuste jaotust mitte kunagi. Tavaliselt peame me üldkogumit kirjeldama vaid ühe valimi arvkarakteristiku põhjal. Hinnates üldkogumi keskväärtust ühe valimi keskväärtuse põhjal teeme me kindlasti mingi vea.
Kuid . . . kas on tõenäolisem, et me teeme väikese või suure vea? (Vt. ülaltoodud joonist!)
* * *

Palju tõenäolisem on, et me teeme väikese vea. Jooniselt on näha, et kõige rohkem on selliseid valimeid, mille keskväärtus erineb üldkogumi keskväärtusest vaid natuke; mida suurem on erinevus, seda vähem on selliseid valimeid ning seega on väiksem tõenäosus saada valimit, mille keskväärtus üldkogumi keskväärtusest palju erineb.

Vaatame nüüd valimite keskväärtuste hajuvust (hajuvus väljendab jaotuse liikmete omavahelist erinevust). Samuti nagu valimite keskväärtuste jaotusel on oma keskväärtus (mis võrdub üldkogumi keskväärtusega) on tal ka oma standardhälve. Kui võrrelda valimite keskväärtuste hajuvust üldkogumi hajuvusega, siis mis te arvate, kas see on suurem, väiksem või enam-vähem sama suur?
* * *

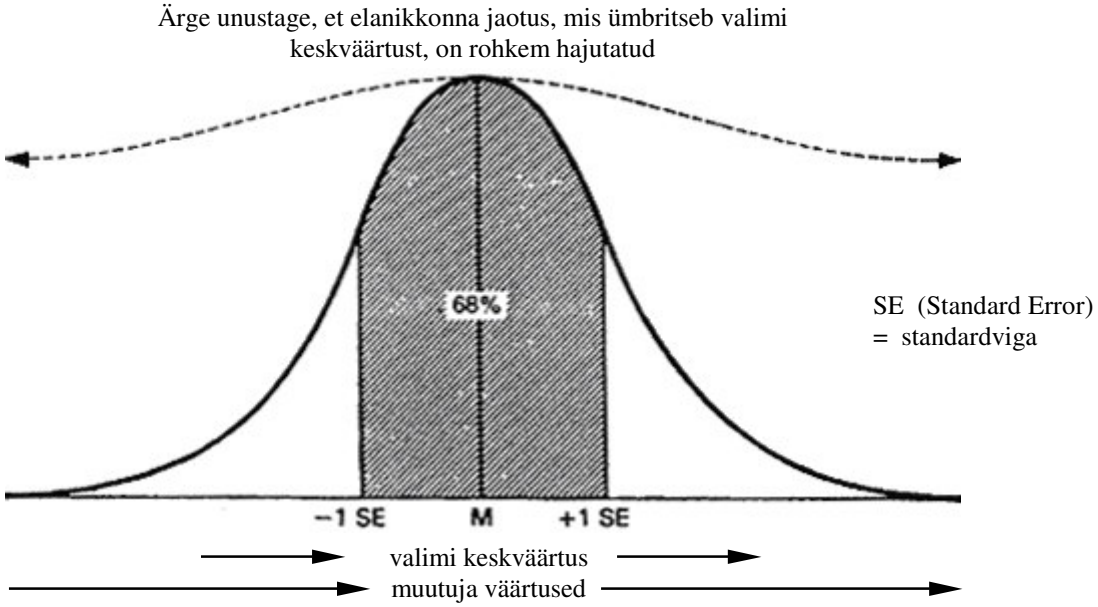
Valimite keskväärtuste jaotuse hajuvus ja seega ka standardhälve on väiksemad kui üldkogumil. Seda illustreerib järgmine joonis, millel on kujutatud üldkogumi jaotus, ühe (küllalt suure) valimi jaotus ning valimite keskväärtuste jaotus:



Milline kõver millist jaotust iseloomustab?
* * *

Üldkogumi jaotusele vastab kõver (A). Valim, mille keskväärtus jääb üldkogumi keskväärtusest veidi paremale, kuid mille ulatus on sama, mis üldkogumil, on (C). Ning valimite keskväärtuste jaotus, mille keskväärtus langeb kokku üldkogumi keskväärtusega kuid mille hajuvus on märgatavalt väiksem nii üldkogumi kui ühe valimi jaotuse hajuvusest, on (B).

Ka siis, kui meil on üldkogumi keskväärtuse hindamiseks kasutada ainult ühe valimi keskväärtus, võime seda valimi keskväärtust vaadelda kui ühte väärtust kõigi *võimalike* valimite keskväärtuste jaotusest. Kui eeldada, et need valimid on küllalt suured ning neid on piisavalt palju, siis vastab see jaotus normaaljaotusele ning me saame kasutada normaaljaotuse omadusi.
Nagu me nägime, on valimite keskväärtuste jaotusel lisaks oma keskväärtusele ka oma standardhälve. Valimite keskväärtuste jaotuse standardhälvet nimetatakse KESKVÄÄRTUSE STANDARDVEAKS. Standardviga lubab meil välja arvutada, kui suur on tõenäosus, et meie poolt moodustatud (ühe) valimi keskväärtus on üldkogumi keskväärtusest märksa suurem või väiksem. (Edaspidi tähistame valimi keskväärtust: V-keskväärtus ning üldkogumi keskväärtust: Ü-keskväärtus)
Kuna valimite keskväärtuste jaotus vastab normaaljaotusele, siis võime öelda, et 68% kõigist valimite keskväärtustest asub vahemikus: Ü-keskväärtus ± 1 st. viga.



See kõik on ilus, kuid tegelikus elus on meie käsutuses siiski ainult ühe valimi keskväärtus. Kuidas me siis võime teada (kõigi võimalike) valimite keskväärtuste jaotuse keskväärtust või standardhälvet?
Matemaatikud on tõestanud, et keskväärtuse standardviga (st. valimite keskväärtuste jaotuse standardhälve) sõltub ainult kolmest faktorist:

1. valimi standardhälbest,
2. valimi suurusest
3. ning sellest, kui suure osa valim üldkogumist hõlmab.

Vaatame neid faktoreid ükshaaval. Kuidas mõjutab valimi standardhälve keskväärtuste jaotuse standardhälvet? Kas valimi suurem hajuvus toob kaasa keskväärtuste jaotuse suurema hajuvuse või vastupidi?
* * *

Mida suurem on erinevus valimite sees, seda suurem on tõenäosus, et selliste valimite keskväärtused erinevad omavahel suurel määral. Seega suurem valimi standardhälve toob kaasa suurema keskväärtuste standardvea.
Kuidas mõjutab aga standardviga valimite suurus? Kas suuremat standardviga võib oodata siis, kui valimid on väikesed või siis, kui nad on suured?
* * *

Mida suuremad on valimid, seda lähemal asuvad nende keskväärtused üldkogumi keskväärtusele. Seepärast, seda väiksem on standardviga.
Vaatame nüüd kolmandat faktorit. Kuidas mõjutab standardviga valimi poolt hõlmatud üldkogumi osa suurendamine?
* * *

Mida suurema osa hõlmab valim üldkogumist, seda väiksem on standardviga. Kuid see faktor on märkimisväärselt *ebaoluline*. Ta omab mingit, kuid väga väikest mõju standardveale. Seepärast jäetakse ta standardvea leidmisel kõrvale.
Tegelikult saadakse keskväärtuse standardviga valimi standardhälbe jagamisel ruutjuurega valimi liikmete arvust. Seega valim, mis koosneb 100-st eksamitulemusest ning mille standardhälve on näiteks 15 punkti, ütleb meile, et kõigi selliste valimite keskväärtuste standardviga on:

$$\frac{15}{\sqrt{100}} = \frac{15}{10} = 1{,}5\,palli = st.viga$$

Selleks, et vähendada standardviga, peaksime me suurendama valimi liikmete arvu (näiteks sajalt neljasajale):

$$\frac{15}{\sqrt{400}} = \frac{15}{20} = 0.75\,palli = st.viga$$

Mida väiksem standardviga, seda kindlamad me võime olla, et meie valimi keskväärtus on lähedane üldkogumi keskväärtusele. Kuid suur valimi liikmete arvu kasv toob kaasa suhteliselt väikese muutuse standardveas. Seepärast tuleb leida optimaalne valimi suurus, mis ühelt poolt annaks piisavalt täpse tulemuse, kuid teiselt poolt oleks uurimuse läbiviimise seisukohast reaalne.

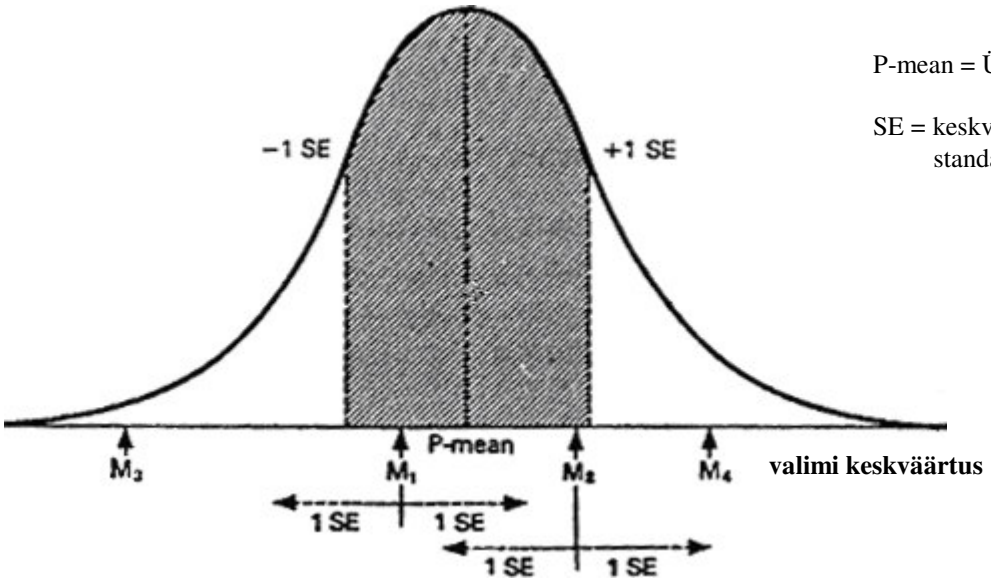
5.4 Üldkogumi keskväärtuse hindamine

Nägime, et keskväärtuse standardviga on määratud valimi suuruse ning standardhälbega. Seega võime öelda, et näiteks vahemik:

$$\text{Ü - keskvä ä rtus} \pm 1\,\frac{\text{valimi st.hä lve}}{\sqrt{\text{valimi suurus}}}$$

ehk Ü-keskväärtus ± 1 st. viga

sisaldab umbes 68% kõigist V-keskväärtustest. Kuid, mis kasu on sellest teadmisest kui me ei tea üldkogumi keskväärtust; veelgi enam - seda me just leida tahamegi. Kõik, mis me teame on ühe valimi keskväärtus ning keskväärtuse standardviga. Kohe vaatame kuidas selle informatsiooni põhjal on võimalik hinnata üldkogumi keskväärtust. Allpool toodud joonisel on kujutatud valimite keskväärtuste jaotus. Märgitud on piirkond, mis jääb Ü-keskväärtusest 1 standardvea võrra kummalegi poole. (Tuletage meelde; see piirkond sisaldab 68% kõigist valimi keskväärtustest!)



Jooniselt on näha, et iga V-keskväärtuse korral, mis jääb märgitud piirkonda (joonisel tähistatud M_1 ja M_2) sisaldab vahemik: **V-keskväärtus ± 1 st. viga**, üldkogumi kesk-väärtust (joonisel P-mean). Kuidas on aga valimitega, mille keskväärtus on näiteks M_3 ja M_4?
* * *

Kuna M_3 ja M_4 asuvad Ü-keskväärtusest kaugemal kui 1 standardviga, siis vahemik, mis jääb sellistest V-keskväärtustest 1 standardvea võrra kummalegi poole, üldkogumi keskväärtust *ei* sisalda (proovi joonistada!).
Kuna me teame, et 68% kõigist võimalikest V-keskväärtustest asub vahemikus: Ü-keskväärtus ± 1 st. viga, siis mistahes valimi korral võime me öelda, et tõenäosusega **68%** asub üldkogumi keskväärtus vahemikus:

V-keskväärtus ± 1 st. viga

Loomulikult ei tohi unustada, et tõenäosusega 32% ta seal ei asu!

Vaatame näidet, kus oli mõõdetud 100 tudengi testitulemused: keskmine tulemus on 50 palli ning standardhälve 15 palli. Seega keskväärtuse standardviga on

$$\frac{15}{\sqrt{100}} = \frac{15}{10} = 1{,}5\,palli$$

Kuidas saaksime nüüd hinnata kõigi testil osalenud tudengite keskmist tulemust? Jah, võime öelda, et see keskmine tulemus asub 68%-lise tõenäosusega vahemikus:

V-keskväärtus ± 1 st. viga ehk

50 ± 1,5 palli ehk 48,5 kuni 51,5 palli.

Seda piirkonda nimetatakse 68 %-liseks USALDUSINTERVALLIKS. Tõenäosust, millega üldkogumi keskväärtus usaldusintervallis asub nimetatakse USALDUSNIVOOKS ning usaldusintervalli otspunkte nimetatakse USALDUSPIIRIDEKS.
Kas sa oskaksid leida piirkonna, milles üldkogumi keskväärtus asub 95%-lise (99%-lise) tõenäosusega? Statistika 'keeles' kõlaks see ülesanne järgmiselt: leia keskväärtuse usaldusintervall usaldusnivool 95% (või 99%).

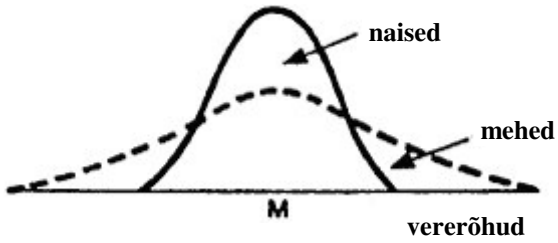
5.5 Teiste üldkogumi parameetrite hindamine.

6. Valimite võrdlemine.

Selles peatükis vaatleme teist järeldava statistika väga tähtsat valdkonda: vaatame kahte erinevat valimit ning küsime, kas erinevus valimi arvkarakteristikute vahel lubab väita, et need valimid kuuluvad erinevatesse üldkogumitesse. Seda laadi mõttekäik peitub sellistes küsimustes nagu: 'Kas tüdrukud on intelligentsemad kui poisid?', 'Kas uus õpetamismeetod annab paremaid tulemusi kui vana', 'Kas nooremate inimeste arvamus erineb vanemate inimeste arvamusest?' jne.

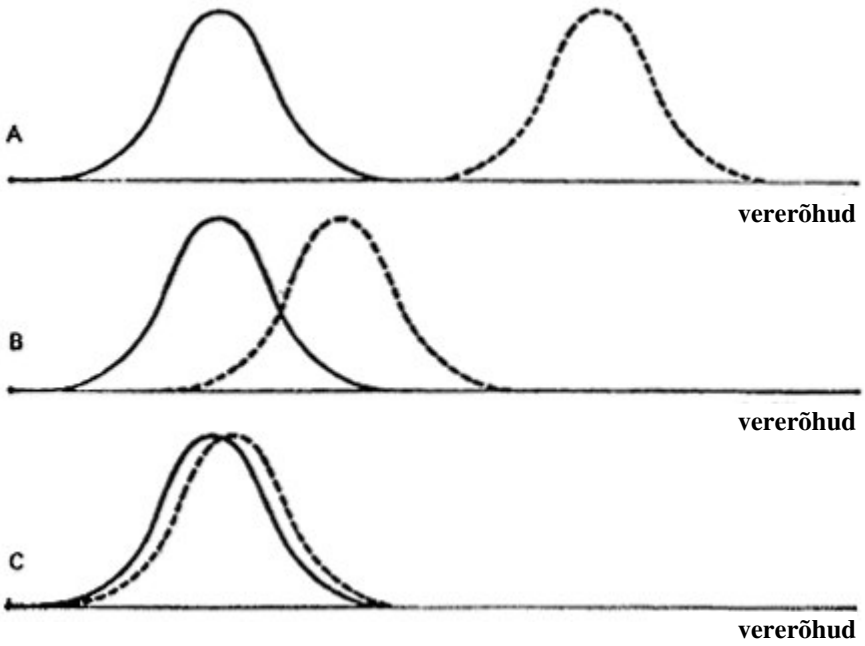
6.1 Kaks valimit: kas samast või erinevatest üldkogumitest? T-test.

Oletame, et me mõõdame katseisikute vererõhku kahes juhuslikus valimis, millest ühte kuulub 50 meest ning teise 50 naist. Millise järelduse me võime nende valimite põhjal teha naiste ja meeste vererõhkude *erinevuse* kohta *üldiselt*. Kas valimid (nende jaotused) on nii sarnased, et me võime öelda, et nad kuuluvad ühte üldkogumisse; või on nad nii erinevad, et esindavad kaht erinevat üldkogumit?
Oletame, et meie valimite jaotuskõverad näevad välja umbes nii:



Kas me võime järeldada, et need valimid kuuluvad ühte üldkogumisse? See tähendab, kas naiste vererõhkude üldkogumi keskväärtus ja standardhälve on samad, mis meeste vererõhkude üldkogumil? Miks?
* * *

Kuigi keskväärtused on valimites peaaegu samad, erinevad standardhälbed omavahel märgatavalt. Seepärast tundub, et need valimid kuuluvad erinevatesse üldkogumitesse. Meeste vererõhud erinevad omavahel märksa rohkem kui naiste vererõhud.
Teine võimalus on selline, kus valimite standardhälbed on küllalt sarnased, kuid *keskväärtused* erinevad üksteisest märgatavalt. Vaata alljärgnevat joonist: sellel on näha kolm erinevat valimite paari. Millisel juhul (A, B või C) võite te kõige kindlamalt väita, et (1) need valimid on erinevatest üldkogumitest ning
(2) need valimid on samast üldkogumist?
* * *



Igas valimite paaris erinevad keskväärtused teineteisest. Juhul (A) on nad väga erinevad, juhul (B) palju vähem erinevad ning juhul (C) on erinevus vaevu märgatav. Meile on teada, et võttes ühest ja samast üldkogumist erinevaid juhuslikke valimeid, ei saa me oodata, et nende valimite keskväärtused oleks täpselt ühesugused. Kuid samuti teame me, et on märksa tõenäolisem saada valimite paar, mille keskväärtused erinevad üksteisest vähesel määral, kui selline paar, mille keskväärtused on väga erinevad. Seepärast võime me väita, et kõige kindlamalt on erinevatest üldkogumitest valimid joonisel (A) ning kõige kindlamalt on ühest üldkogumist valimid juhul (C).

Niisiis, võttes kaks erinevat juhuslikku valimit, saame me alati mingil määral erinevad keskväärtused. Kui suur see erinevus peaks aga olema, et me võiksime teha järelduse, et meie valimid esindavad erinevaid üldkogumeid. Ehk teisisõnu: kui palju peavad erinema näiteks katses osalenud tüdrukute ja poiste keskmised testitulemused, et me võiksime väita et tüdrukud *üldiselt* on poistest aktiivsemad või passiivsemad, targemad või rumalamad, rahulikumad või kergemini ärrituvad jne.

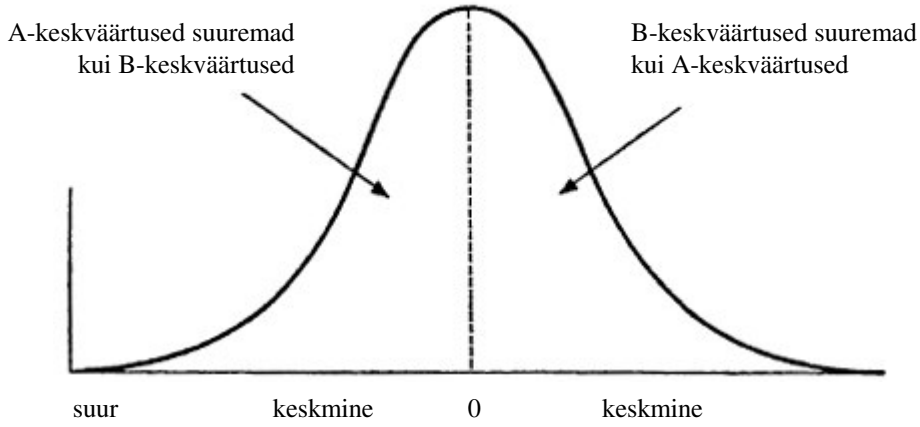
6.2 Olulisuse testid

Valimite sellist võrdlemist nimetatakse OLULISUSE TESTIMISEKS. Me kontrollime, kas erinevus valimite vahel on piisav tõestamaks tegelikku erinevust üldkogumite vahel. Edaspidi vaatamegi kuidas sellist olulisustesti tehakse. Olgu vahemärkusena lisatud, et olulisustesti läbiviimiseks on mitu võimalust. Siinkohal käsitleme vaid statistikas kõige enam kasutatavat meetodi, mis põhineb valimite keskväärtuste vaheliste ERINEVUSTE JAOTUSEL. See idee võib esmapilgul nõuda teilt veelgi enam kujutlusvõimet, kui valimite keskväärtuste jaotuse mõistmine, kuid ma loodan, et te harjute selle mõtteviisiga peagi.

Kujutage ette olukorda, kus me võtame **ühest üldkogumist** näiteks sada juhuslikku valimit. Sel korral võtame aga korraga kaks valimit: A ja B. Loomulikult võime me saada sellised valimid, mille keskväärtused on täpselt ühesugused, kuid sagedamini saame me sellised valimite paarid, mille keskväärtused üksteisest veidike erinevad. Võttes kirjeldatud viisil lõpmatult palju valimite paare, saaksime moodustada jaotuse valimite *keskväärtuste erinevustest.* Teisisõnu, me küsime: kui palju esineb selliseid valimite paare, mille puhul A keskväärtus on palju suurem kui B

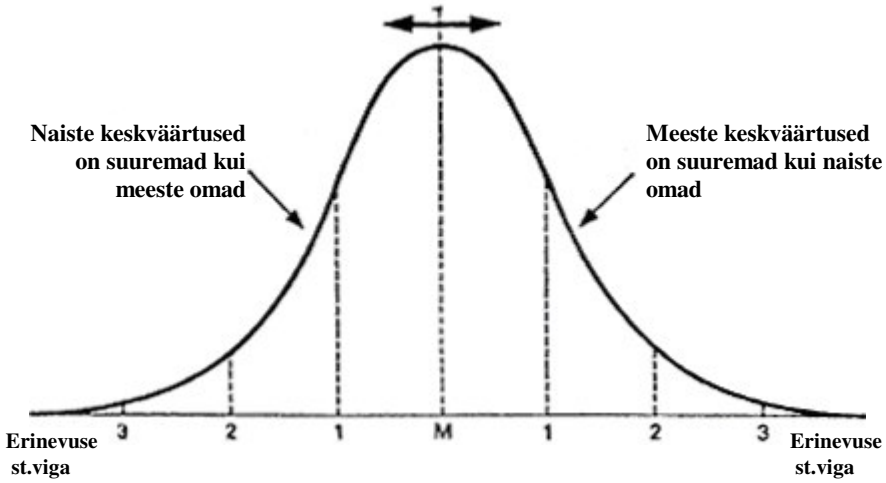
keskväärtus? Kui sageli on A keskväärtus natuke suurem kui B keskväärtus? Kui sageli on valimite keskväärtused võrdsed? Kui sageli on B keskväärtus natuke suurem kui A keskväärtus? jne.
Kas te oskate juba ette kujutada, milline võiks välja näha see valimi keskväärtuste erinevuste jaotus kui me võtaks ühest üldkogumist väga palju juhuslike valimite paare?
* * *

See valimite paaride keskväärtuste *erinevuste* jaotus vastaks *normaaljaotusele*. Selle jaotuse keskväärtus oleks null (see vastab olukorrale, kus kahe valimi keskväärtused üksteisest ei erine). A-keskväärtus oleks B-keskväärtusest suurem sama sageli kui B-keskväärtus oleks suurem A-keskväärtusest ning väiksem erinevus kahe valimi keskväärtuste vahel esineks sagedamini kui suurem. Jaotuskõver näeks siis välja umbes selline:



Erinevus valimite keskväärtuste vahel

Saadud jaotus näitab, millist erinevust võib oodata kahe juhuslikult valitud valimi korral, mis on võetud ühest üldkogumist.
Nagu igal jaotusel, nii on ka valimite keskväärtuste erinevuste jaotusel oma keskväärtus ning standardhälve. Nagu juba nägime on sellise jaotuse keskväärtuseks null (mis väljendab seda, et erinevust kahe keskväärtuse vahel ei ole). Erinevuste jaotuse standardhälve sõltub aga üldkogumi standardhälbest ning on seda suurem, mida suurem on üldkogumi standardhälve. Samuti nagu valimi keskväärtuste jaotuse puhul nimetatakse ka valimi keskväärtuste erinevuste jaotuse standardhälvet standardveaks - seekord aga vastavalt KESKVÄÄRTUSTE VAHELISE ERINEVUSE STANDARDVEAKS.
Kuna on tegemist normaaljaotusele vastava jaotusega, siis kehtivad siin kõik normaaljaotuse proportsioonid: st. ligikaudu 68% kõigist valimi keskväärtuste vahelistest erinevustest jäävad piirkonda **0 ± 1 erinevuse st. viga**; ligikaudu 95% erinevustest jäävad piirkonda **0 ± 2 erinevuse st. viga** jne.



Selliselt oleme konstrueerinud jällegi ühe abstraktse jaotuse, mida me tegelikus elus vaevalt kunagi kohtame. Kuid tuletage meelde, millisele küsimusele me peatüki alguses vastust otsima hakkasime!

* * *

Tõepoolest, meil oli ju tegemist kahe valimiga ning me tahtsime teada, kas erinevus nende valimite keskväärtuste vahel lubab meil väita, et erinevus esineb ka üldkogumi tasandil ehk teisisõnu, kas erinevus meie valimite keskväärtuste vahel näitab, et need valimid on erinevatest üldkogumitest (st. üldkogumitest, mille keskväärtused üksteisest erinevad) või on see erinevus tingitud lihtsalt juhusest.

Vaatame ühte näidet. Olgu meil mõõdetud 50 meessoost ning 50 naissoost tudengi vererõhud. Meeste keskmine vererõhk oli 120 mm ning naiste keskmine vererõhk oli 110 mm. Standardhälve oli mõlema jaotuse puhul 11,3 mm.
Seega erinevus valimite keskväärtuste vahel on 10 mm. Kas selline erinevus lubab meil väita, et nais- ja meessoost tudengite vererõhud ka üldiselt erinevad st. kui meil oleks võimalik mõõta kõigi nais- ja meessoost tudengite vererõhku kas siis nende keskväärtuste vahel esineks samuti erinevus? Selle otsuse tegemiseks peame me läbi viima olulisuse testi. Kuid - statistikud ja ka teised teadlased on ettevaatlik rahvas. Seepärast ei ürita nad tõestada, et erinevus, mis on saadud valimite keskväärtuste vahel on oluline (näitab tegelikku erinevust üldkogumites) vaid seavad üles küsimuse : “Kas on võimalik, et saadud erinevus pole oluline?”
Seega alustame me oletusega, et tegelikkuses ei esine erinevust naiste ja meeste vererõhkude vahel. Me eeldame, et nad kõik on ühest üldkogumist ning erinevus, mis me valimite vahel saime on lihtsalt üks juhuslik väärtus eespool kirjeldatud valimi keskväärtuste erinevuste jaotusest. Sellist väidet nimetatakse statistikas NULLHÜPOTEESIKS.
Siitpeale tegeleme nullhüpoteesiga. Kui erinevus kahe valimi keskväärtuste vahel on liiga suur selleks, et seda saaks pidada erinevuseks, mis sageli saadakse kahe ühest üldkogumist võetud juhusliku valimi vahel, siis tuleb nullhüpotees ümber lükata. Ümberlükatud nullhüpotees tuleb nüüd asendada uue ALTERNATIIVSE HÜPOTEESIGA. Kõige sagedamini on alternatiivseks hüpoteesiks väide, et üldkogumite keskväärtused *ei ole võrdsed* (st. erinevus on oluline).
Millised võiks veel olla alternatiivsed hüpoteesid?

* * *

Alternatiivseks hüpoteesiks võiks veel olla näiteks väide, et meeste vererõhk on kõrgem kui naistel või ka vastupidi.

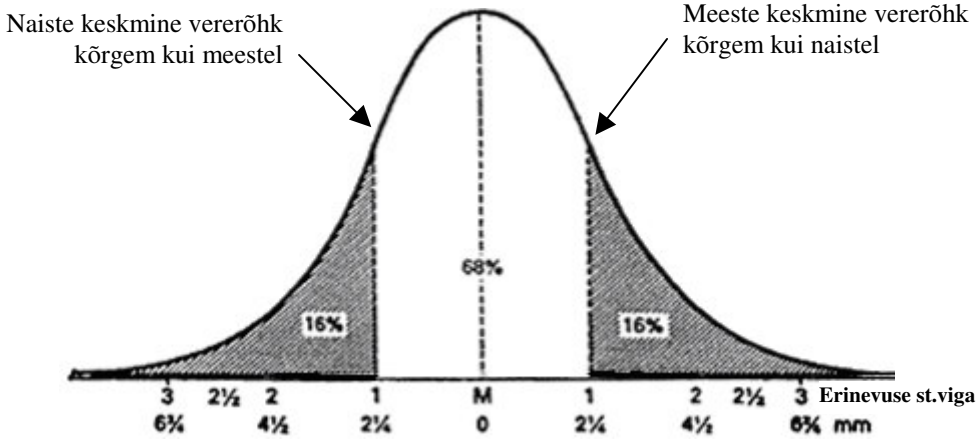
Vaatamata sellele, milline on alternatiivne hüpotees, jääb nullhüpotees kehtima seni kuni me pole suutnud näidata, et erinevus valimite vahel on liiga suur selleks, et need valimid võiksid olla ühest üldkogumist. Niisiis oletame me, et erinevus 10 mm on lihtsalt üks võimalikest erinevustest, mis kuuluvad ühest üldkogumist võetud valimipaaride keskväärtuste erinevuste jaotusesse (vt. ülal toodud joonist).
Selle jaotuse keskväärtus on null. Kuid kuidas leida selle jaotuse standardhälvet e. keskväärtuste erinevuse standardviga? Loomulikult ei saa me leida tegelikku standardhälvet võtmata üldkogumist lõpmatut hulka valimite paare. Seepärast tuleb meil erinevuse standardviga *hinnata*. Matemaatikud on tõestanud, et keskväärtuste vahelise erinevuse standardviga saab leida kombineerides mõlema valimi keskväärtuste standardvead.
Meie näites on kummagi viiekümneliikmelise valimi standardhälve 11,3 mm. Seepärast on ka keskväärtuse standardviga mõlemal juhul sama:

$$\frac{11,3}{\sqrt{50}} = 1,6mm$$

Erinevuste standardvea saame liites kokku valimite standardvigade ruudud ning võttes saadud summast ruutjuure:

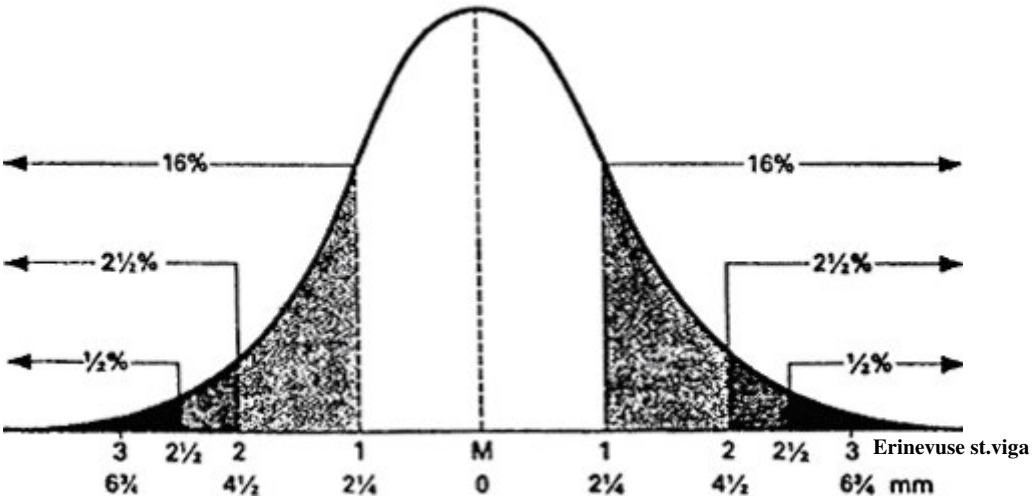
$$erinevuse\_st.viga = \sqrt{1,6^2 + 1,6^2} = \sqrt{2.56 + 2,56} = \sqrt{5,12} = 2,26mm$$

Nüüd saame konstrueerida ühest üldkogumist võetud juhuslike valimite keskväärtuste erinevuste jaotuse:



Valimite keskväärtuste vaheline erinevus

Jooniselt on näha, et *juhul, kui tegelikkuses ei esineks mingit erinevust* naiste ja meeste vererõhkude vahel oleks 100-st valimite paarist 68-l juhul keskväärtuste erinevus väiksem kui 2,26 mm. Juhul kui meie naiste ja meeste valimi keskväärtuste erinevus oleks olnud 2,26 millimeetrist väiksem, siis oleksime järeldanud, et see erinevus on tingitud juhusest ning on liiga väike tõestamaks tegelikku erinevust naiste ja meeste vererõhkude vahel.
Tegelikult oli aga meie näites valimite erinevus 10 mm. Kus paikneb see erinevus kõigi võimalike erinevuste jaotuses?



Valimite keskväärtuste vaheline erinevus

Näeme, et meie erinevus on suurem kui ±3 erinevuse st. viga. Kuna kõikvõimalike erinevuste jaotus vastab normaaljaotusele, siis on teada, et ühest üldkogumist võetud juhuslike valimite paaride puhul on 99-l juhul 100-st erinevus väiksem kui ± 2,5 erinevuse st. viga. Seega on tõenäosus saada ühest üldkogumist nii suure erinevusega juhuslike valimite paar alla ühe protsendi. Seepärast tuleb nullhüpotees (mis väitis, et erinevus pole oluline) ümber lükata ning asendada alternatiivse hüpoteesiga, mis väidab, et erinevus valimite keskväärtuste vahel on oluline. Teisisõnu: saadud erinevus viiekümne meessoost ja viiekümne naissoost tudengi keskmiste vererõhkude vahel lubab meil väita, et meessoost ja naissoost tudengite keskmiste vererõhkude vahel üldiselt on olemas erinevus.

Nüüd jääb üle vaid küsimus, kui väike peab olema tõenäosus saada ühest üldkogumist antud erinevusega valimite paar (seda tõenäosust nimetatakse OLULISUS-TÕENÄOSUSEKS ning tähistatakse tavaliselt p), et me võiksime nullhüpoteesi ümber lükata? Tegelikult peab selle piiri valima iga uurija ise ja seda enne andmete kogumist-analüüsimist.
Statistikas on saanud traditsiooniks kasutada OLULISUSNIVOOSID (tähistatakse α) 0,01 (ehk 1%) ja 0,05 (ehk 5%). Valides olulisusnivooks 0,05 peab olulisustõenäosus selleks, et nullhüpoteesi ümber lükata olema väiksem kui 0,05 ning vastavalt olulisusnivoo 0,01 korral peab ta olema väiksem kui 0,01.

6.3 Ühe- ja kahepoolsed testid

Eelmises punktis vaatlesime juhtu, kus meil on vaja kontrollida, kas erinevus kahe valimi keskväärtuse vahel on statistiliselt oluline st kas me võime antud valimite põhjal mingil etteantud olulisusnivool väita, et üldkogumite keksväärtused on erinevad:

tõestatav hüpotees: $H_1: \mu_1 \neq \mu_2$ (erinevus valimite keskväärtuste vahel on oluline e üldkogumite keskväärtused on erinevad)

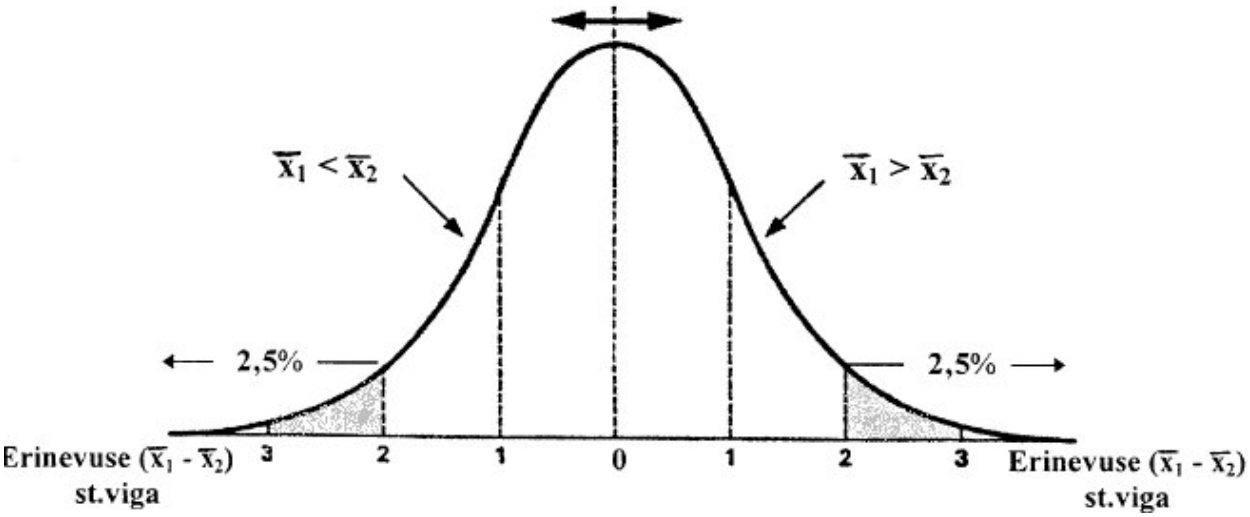
nullhüpotees: $H_0: \mu_1 = \mu_2$ (erinevus valimite keskväärtuste vahel ei ole oluline e üldkogumite keskväärtused ei ole erinevad)

Tihtipeale on meil aga soov tõestada, et ühe üldkogumi keskväärtus on teisest suurem või kõrgem (mitte lihtsalt erinev). Näiteks soovime me näidata, et uus õppemeetod annab parema tulemuse (kõrgema keskmise testitulemuse kursuse lõpus) kui vana õppemeetod või et kuuenda klassi tüdrukute keskmine pikkus on suurem kui poistel jne. Seega on nüüd

tõestatavaks hüpoteesiks: $H_1: \mu_1 < \mu_2$ või $\mu_1 > \mu_2$

Tuletame meelde, millised sammud tuli astuda, et hüpoteesi $H_1: \mu_1 \neq \mu_2$ tõestada.

1. Püstitada nullhüpotees ($H_0: \mu_1 = \mu_2$)
2. Valida olulisusnivoo (näiteks 0,05 e 5%)
3. Konstrueerida kõikvõimalike juhuslike (valimite keskväärtuste vaheliste) erinevuste jaotus ning leida kriitilised väärtused vastavalt valitud olulisusnivoole.



4. Võrrelda valimite keskväärtuste vahel saadud erinevust ($\overline{x}_1 - \overline{x}_2$) konstrueeritud jaotusega. (Kui suur on tõenäosus, et saadud erinevus on juhuslik?)
5. Kui (olulisus)tõenäosus on suurem kui olulisusnivoo, siis peame jääma nullhüpoteesi juurde ning seega hüpoteesi $H_1: \mu_1 \neq \mu_2$ ei ole õnnestunud tõestada;
   kui (olulisus)tõenäosus on väiksem kui olulisusnivoo, siis võib nullhüpoteesi ümber lükata ning seega on hüpotees $H_1: \mu_1 \neq \mu_2$ tõestatud.

Sellist olulisustesti, kus me ei eelda, et üks keskmine on teisest kindlasti suurem vaid vaatleme nii juhtu $\overline{x}_1 > \overline{x}_2$ kui juhtu $\overline{x}_1 < \overline{x}_2$ nimetatakse **kahepoolseks olulisustestiks** (vt ülalolevat joonist).

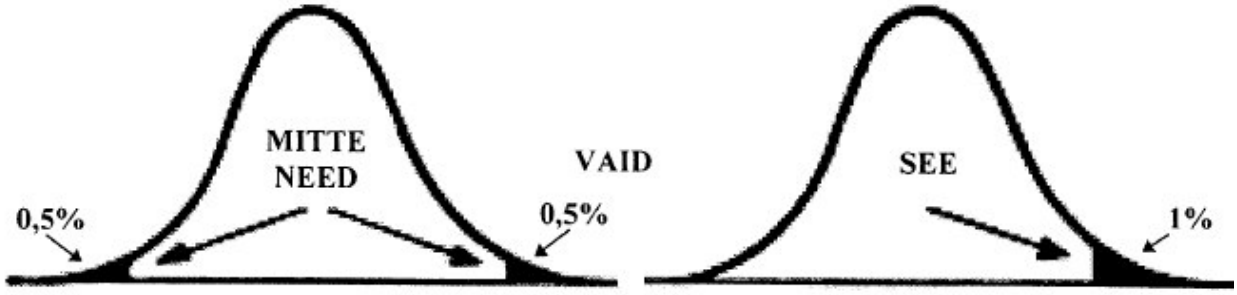
Kui tõestatavaks hüpoteesiks on aga H_1: $\mu_1 < \mu_2$ või H_1: $\mu_1 > \mu_2$, siis me eeldame, et üks keskmine on teisest kindlasti suurem ning võiksime kasutada **ühepoolset olulisustest.**

Võtame ühe konkreetse näite:
Katsetatakse uut õppemetoodikat ning soovitakse näidata, et ta on efektiivsem kui vana. Seetõttu moodustatakse kaks katsegruppi, milles on mõlemas 100 õpilast. Ühte gruppi õpetatakse uue teist vana metoodika järgi ning kursuse lõpul viiakse mõlemas grupis läbi kontrolltest. Katsetulemuste objektiivsuse tagamiseks on korraldatud nii, et õpilased ei tea kumma metoodika järgi neid õpetatakse.

Tõestatavaks hüpoteesiks on meil siis H_1: $\mu_u > \mu_v$
1. Püstitame nullhüpoteesi H_0: $\mu_u \leq \mu_v$
2. Valime olulisusnivooks 0,01 e 1%

Seega selleks, et nullhüpoteesi ümber lükata, peab uue metoodika järgi õpetatud õpilaste keskmine testitulemus $\bar{x}_u$ olema vana metoodika järgi õpetatud õpilaste keskmisest testitulemusest $\bar{x}_v$ nii palju **kõrgem**, et meil jääks ainult üks võimalus sajast, et see erinevus on juhuslik.

Vaatame jälle kõikvõimalike juhuslike erinevuste jaotust, kuid nüüd rahuldavad meid ainult ühel pool olevad erinevused - need, mille puhul $\bar{x}_u > \bar{x}_v$ seepärast ei peaks me
vaatama mitte sellist joonist: vaid hoopis sellist joonist:



Kui suur peab olema erinevus uue metoodika kasuks, et tõenäosus teda juhusikult saada oleks väiksem kui 1%?
a) 2,5 erinevuse standardviga
b) võib olla väiksem kui 2,5 erinevuse standardviga
c) peab olema suurem kui 2,5 erinevuse standardviga
***

Erinevus võib olla väiksem kui 2,5 erinevuse standardviga, sest on teada, et tõenäosus saada erinevust mis on 2,5 standardviga on ainult 0,5% (ning mida suurem erinevus, seda väiksem tõenäosus on teda juhuslikult saada)

Ligikaudsete arvutuste kohaselt on ühepoolse olulisustesti kriitiliseks piiriks olulisusnivool 1% 2,33 standardviga ning
olulisusnivool 5% 1,67 standardviga.

Seega nägime, et ühepoolse hüpoteesi tõestamiseks peab erinevus valimi keskväärtuste vahel olema väiksem kui kahepoolse hüpoteesi tõestamiseks (ehk ühepoolset hüpoteesi on lihtsam tõestada). Seepärast peab ühepoolse hüpoteesi kasutamine olema teoreetiliselt väga hästi

põhjendatud (millised teoreetilised eeldused on mul väita, et just uus metoodika annab kõrgema testitulemuse ja mitte vana?). Tihti kasutatakse ka ühepoolse hüpoteesi tõestamiseks kahepoolset olulisustesti, sest kui õnnestub nullhüpotees ümber lükata kahepoolse olulisustesti korral, siis võib ta kindlasti ümber lükata ka ühepoolse olulisustesti korral. Loomulikult tuleb niisugusel juhul enne alternatiivse hüpoteesi tõestatuks lugemist veenduda, et erinevus valimite keskväärtuste vahel on ikka teie hüpoteesiga samasuunaline (kas $\overline{x}_u > \overline{x}_v$).

ANDMED:

	$\overline{x}$	st. hälve
uus metoodika	62,8 p	10 p
vana metoodika	60 p	9 p
erinevus:	**2,8 p**	

$$\text{st. viga}_{(\overline{x}_u)} = \frac{10}{\sqrt{100}} = 1p$$

$$\text{st. viga}_{(\overline{x}_v)} = \frac{9}{\sqrt{100}} = 0{,}9p$$

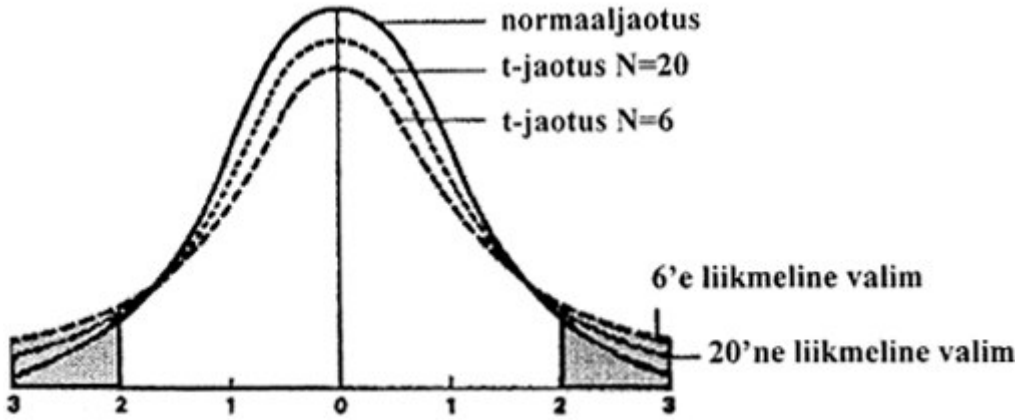
$$\text{erinevuse st. viga}_{(\overline{x}_u - \overline{x}_v)} = \sqrt{1^2 + 0{,}9^2} = \sqrt{1{,}81} \approx 1{,}3p$$

Kriitilise piiri olulisusnivool 1% saame korrutades 2,33*1,3p ≈ 3,0p

Meie erinevus (2,8) on väiksem kui saadud kriitiline väärtus (3,0) seetõttu peame jääma nullhüpoteesi juurde: meil ei õnnestunud olulisusnivool 1% tõestada, et uus metoodika annab kõrgema testitulemuse kui vana.

6.4 z-testid ja t-testid

Eelmistes punktides vaadeldud olulisusteste võiks nimetada z-testideks, kuna me kasutasime standardhälvet kui ühikut ning kriitilised väärtused leidsime kogu aeg normaaljaotuse proportsioone silmas pidades (ligikaudseid küll, aga siiski).
Matemaatik William Gossett varjunimega 'Student' märkas aga, et väikeste valimite korral on valimi standardhälve tavaliselt suurem kui üldkogumi oma ning ühtlasi, et mida väiksem valim, seda suurem on erinevus üldkogumi standardhälbest. Seepärast tuleks kriitiliste väärtuste leidmisel kasutada mitte normaaljaotuse vaid t-jaotuse proportsioone.
t-jaotus on normaaljaotusega väga sarnane; ainus erinevus on selles, et t-jaotuse hälve on suurem (jaotuskõver on laiem) ning sõltub valimi suurusest. Seega ei esinda t-jaotust mitte üks kõver vaid kõverate parv. Järgmisel joonisel on toodud võrdlemiseks kolm jaotuskõverat: normaaljaotus, t-jaotus 20ne liikmega valimi korral ning t-jaotus 6e liikmega valimi korral. Jooniselt on hästi näha, et kriitiline väärtus, mis eraldab samasuure osa (n 5%) ekstreemseid väärtusi peab t-jaotuse puhul olema suurem kui normaaljaotuse puhul (võrdle pindalasid märgitud piirkonnas).



t-väärtus näitab siis erinevuse kohta t-jaotuses ning statistikaprogrammide poolt arvutatav olulisustõenäosus (*Significance*) näitab, kui suur tõenäosus on saada nii suurt erinevust valimi keskväärtuste vahel juhuse tõttu kui üldkogumid (st nende keskväärtused) ei erine.

6.5 I ja II tüüpi viga

Tuletan siinkohal veelkord meelde, et kuitahes suuri pingutusi me ka ei teeks 100%list kindlust me oma järeldustes statistika abil ei saavuta. Tihti öeldakse, et võttes olulisusnivooks 0,05 jätame omale võimaluse eksida viiel juhul sajast, võttes aga olulisusnivooks 0,01 jätame omale võimaluse eksida ühel juhul sajast. See on küll mõnes mõttes õige, kuid me peaks põhjalikumalt mõtlema eksimise võimaluste üle.
Kui meil õnnestub näidata, et erinevus meie valimite keskväärtuste vahel, võib olla juhuslik tõenäosusega alla 5% (või alla 1%) siis võime nullhüpoteesi ümber lükata ning väita, et erinevus on olemas ka üldkogumite keskväärtuste vahel. Seejuures teame, et meil on siiski väike võimalus, et erinevus oli saadud juhuslikult olukorrast, kus üldkogumid tegelikult ei erine. **Sellist viga, kus meie väidame, et üldkogumid on erinevad aga tegelikult nad seda ei ole, nimetatakse I tüüpi veaks ning selle vea suurust reguleerib olulisusnivoo.**
Niisiis tekib esimest tüüpi viga siis, kui nullhüpotees on õige aga meie väidame, et ta on vale. Aga võib ju olla ka vastupidi: nullhüpotees on tegelikult vale aga meie jätame ta ümber lükkamata st. me jätame märkamata erinevuse, mis tegelikult üldkogumite vahel valitseb (vt joonist):

		H_0	
		õige	**vale**
Meie otsus	**ümber lükata**	**I tüüpi viga**	**-**
	vastu võtta	**-**	**II tüüpi viga**

Sel juhul teeme teist tüüpi vea. Kui suur võiks olla võimalus eksida niipidi?
***
Selleks, et vaadata kas meie erinevus sobib kõikvõimalike juhuslike erinevuste hulka küsisime: kui suur on tõenäosus, et meie erinevus on juhuslik? ning sellele andis vastuse olulisustõenäosuse näitaja α (S*ignificance*). Kui α võrduks 0,084ga, siis võiksime öelda, et tõenäosus saada valimite vahel nii suurt erinevust juhuse tõttu on 8,4% ning peaksime jääma nullhüpoteesi juurde. Kuid! tõenäosus et see erinevus ei ole juhuslik on ju 1-α ehk 100%-

8,4%=91,6%. Seega on meil antud juhul võimalus teha II tüüpi viga tõenäosusega 91,6%! Seepärast ei loeta nullhüpoteesi kunaga tõestatuks vaid oma järelduses tõdetakse lihtsalt, et valitud olulisusnivool ei õnnestunud alternatiivset hüpoteesi tõestada.
Seda, kumba vea me tegelikult teeme, ei saa me kunagi teada, küll on aga ilmne, et mida rohkem me püüame hoiduda tegemast esimest tüüpi viga, seda suurem on teist tüüpi vea tegemise oht. Seepärast tuleb enne olulisusnivoo valimist põhjalikult kaaluda, kuivõrd oluline on meil hoiduda esimest tüüpi veast (mis võivad olla selle vea tagajärjed).

Lõpetuseks tahaksin rõhutada, et statistiliselt oluliseks osutunud erinevus ei pruugi elulises mõttes olla sugugi suur ega tähtis ega huvitav erinevus. Isegi kui meil õnnestub olulisusnivool 1% tõestada, et uus õppemetoodika annab kõrgema testitulemuse (ehk paremad teadmised) kui vana, siis ei ole see piisav väitmaks, et kõik õpetajad peavad või peaksid hakkama kasutama uut metoodikat. Enne sellist otsust tuleb vaagida veel palju küsimusi: kui suuri kulutusi uue õppemetoodika rakendamine nõuab, kas õpetajate ettevalmistus ja ka motivatsioon on üleminekuks piisav, kas leidub ehk mõni teine uus metoodika, mis annab samasugused tulemused väiksema kulu ja vaevaga jne. jne.

7. Mitteparameerilised meetodid. χ^2-test.

Klassikalised statistilised meetodid nõuavad tavaliselt, et uuritavad tunnused oleks numbrilised ning nende jaotus vastaks enam-vähem normaaljaotusele. Siit ka nende nimetus: **parameetrilised meetodid**. Tegelikes uurimustes tuleb aga tihtipeale tegelda tunnustega, mis on mõõdetud järjestus- või koguni nominaalskaalal ja/või mille jaotus ei pruugi vastata normaaljaotusele. Siin tulevad meile appi **mitteparameetrilised meetodi**, milles keskväärtuste ja standardhälvete asemel kasutatakse kas järjenumbreid, sagedusi vms.
Üks enam kasutatav viis kategoriaalsetest tunnustest ülevaate saamiseks on ühe- või mitmemõõtmeliste sagedustabelite moodustamine. Kaks tunnust, mis moodustavad kahemõõtmelise sagedustabeli (e risttabeli) võivad olla nii nominaalsed kui ordinaalsed, peaasi, et gruppe ei tekiks liiga palju. Tihtipeale on üks tunnustest teisest sõltuv (n õnnelikkus sõltub soost aga mitte vastupidi), kuid võib olla ka nii, et sõltuvus on vastastikune (n laste arv sõltub haridusest, kuid võib olla ka nii, et haridus sõltub laste arvust).
Mitmed mitteparameetrilised meetodid põhinevadki risttabelil ning lähtuvad selles peegelduva(te)st sagedusjaotus(t)est. Üks enamkasutatavaid meetodeid on kahtlemata **χ^2-test**, mis oma olemuselt on olulisustest ning annab vastuse küsimusele: kas erinevus (kahe) grupi sagedusjaotustes on statistiliselt oluline või mitte e kas erinevus valimite proportsioonides lubab meil väita, et proportsioonid üldkogumites on erinevad.

Võtame ühe lihtsa näite: meie eesmärgiks on uurida immigrantide huvi kohaliku poliitika vastu Tallinnas. Küsitletute arv on 200, neist 120 meest ja 80 naist. Valimi tulemuste põhjal saame järgmise risttabeli:

Huvitub kohalikust poliitikast:	Jah	Ei	
Mees	35	85	120
	≈30%	≈70%	100%
Naine	15	65	80
	≈20%	≈80%	100%
	50	150	
	25%	75%	

Näeme, et valimisse sattunud meessoost immigrantidest on kohalikust poliitikast huvitatud ≈30%, kuid naissoost immigrantidest vaid ≈20%, seega on naiste ja meeste poliitiline huvitatus meie valimis erinev. Edasi peaksime aga küsima, kas see erinevus on piisavalt suur selleks, et me võiksime oma järeldusi üldistada ja väita mingi küllalt suure tõenäosusega, et Tallinna meessoost immigrandid on kohalikust poliitikast enam huvitatud kui naissoost immigrandid (või on see erinevus nii väike, et võib olla tekkinud lihtsalt juhuse tõttu)?
Niisiis oleme jällegi olukorras, kus peaksime oma oletuse kontrollimiseks püstitama nullhüpoteesi, mis võiks kõlada umbes nii: Tallinna mees- ja naissoost immigrantide huvi kohaliku poliitika vastu ei erine e täpsemalt, naistest on kohalikust poliitikast huvitatud sama suur osa kui meestest. Seda (nullhüpoteesi) olukorda nimetatakse tavaliselt oodatud olukorraks. χ^2-testi eesmärgiks ongi võrrelda tegelikku olukorda oodatud olukorraga e tegelikke sagedusi oodatud sagedustega.

Millised võiks olla oodatud proportsioonid, kui mingit erinevust naiste ja meeste vahel ei oleks?
***

Kasutades risttabelis olevaid protsente võib öelda, et oodatavalt peaks nii naistest kui meestest olema kohalikust poliitikast huvitatud 25%. χ^2-test ei võrdle aga protsente vaid tegelikke ja oodatud sagedusi, seepärast tuleks meil kõigepealt arvutada iga lahtri jaoks oodatav sagedus. **Oodatava sageduse leidmiseks tuleb korrutada vastava rea ja vastava veeru kogusagedus ning jagada see katseisikute arvuga.**

Näiteks, esimese lahtri oodatav sagedus on: $\frac{120*50}{200} = 30$

Kontrollime, kas 30 on 25% 120st: $\frac{120*25}{100} = 30$ On!

Teiste lahtrite oodatavad sagedused on siis:

$\frac{120*150}{200} = 90$, $\frac{80*50}{200} = 20$ ja $\frac{80*150}{200} = 60$

Saame risttabeli:

	Huvitub kohalikust poliitikast:	Jah	Ei	
Mees	Tegelik	35	85	120
	Oodatav	*30*	*90*	*120*
Naine	Tegelik	15	65	80
	Oodatav	*20*	*60*	*80*
	Tegelik	50	150	200
	Oodatav	*50*	*150*	

Edasi peaksime vaatama tegeliku ja oodatava sageduse erinevust igas lahtris (T-O). Kui me nüüd kõik need erinevused kokku liidaks siis tuleks vastuseks null, sest oodatavast olukorrast on positiivseid ja negatiivseid erinevusi ühepalju. Seepärast tõstetakse erinevused enne liitmist ruutu ning ka saadavat erinevust iseloomustavat arvnäitajat nimetatakse mitte lihtsalt χ (loe: hii) vaid χ^2.

$$\chi^2 = \sum \frac{(T-O)^2}{O}$$

Kas panete tähele, et enne liitmist on erinevuse ruut veel jagatud oodatava sagedusega! Miks?

***

Vaatame ühte näidet: kui tegelik sagedus lahtris on 61 ja oodatav sagedus on 56, siis erinevus nende vahel on 5. Samuti saame erinevuseks 5, kui meil tegelik sagedus on 9 ja oodatav sagedus 4. Kas see erinevus on aga mõlemal juhul sama tähendusega?

***

Muidugi ei ole! Seepärast leitaksegi erinevuse osakaal oodatava sageduse suhtes (erinevuse ruut jagatakse oodatava sagedusega).

Arvutame nüüd χ^2 väärtuse meie näite jaoks:

$$\chi^2 = \frac{(35-30)^2}{30} + \frac{(85-90)^2}{90} + \frac{(15-20)^2}{20} + \frac{(65-60)^2}{60} = \frac{25}{30} + \frac{25}{90} + \frac{25}{20} + \frac{25}{60} \approx 4{,}17$$

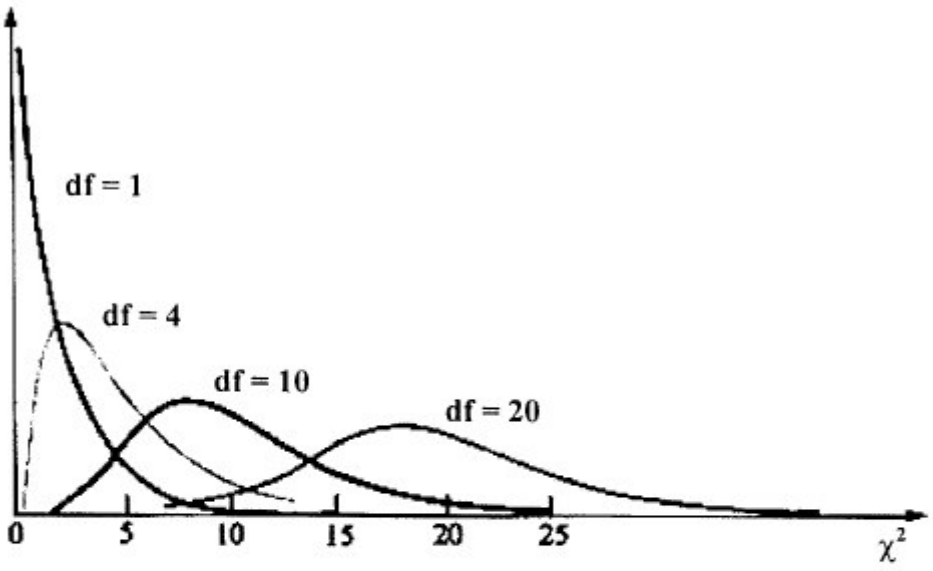
Edasi peaksime saadud χ^2-du väärtust võrdlema kõikvõimalike juhuslike χ^2-de jaotusega ning küsima, kas ja kui suure tõenäosusega meie χ^2 sobib nende juhuslike χ^2-de hulka e kui suur on tõenäosus, et saadud erinevus valimite proportsioonides on tekkinud nullhüpoteesi olukorras (tegelikku erinevust ei ole) juhuse tõttu.

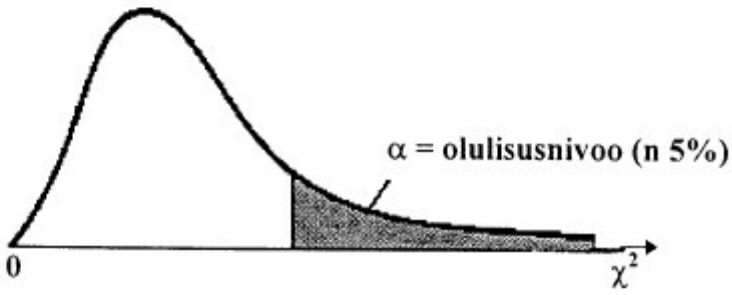
χ^2-du väärtused ei allu kahjuks mingile universaalsele väärtuste skaalale ning on ilmne, et mida rohkem on risttabelis lahtreid, seda suurem võib lihtsalt juhuse tõttu tulla χ^2 -du väärtus. Seetõttu polegi ühte universaalset juhuslike χ^2-de jaotust vaid tegemist on jaotuste perega. Õigete proportsioonidega jaotuse valikul tuleb lähtuda **vabadusastmete arv**ust, mis risttabeli puhul saadakse korrutades tabeli ühe võrra vähendatud ridade arv tabeli ühe võrra vähendatud veergude arvuga:

$$df = (r.arv-1)*(v.arv-1)$$

Meil: df = (2-1)*(2-1) = 1

Kõikvõimalike juhuslike χ^2-de jaotused näevad välja niimoodi:





Meie poolt eelnevalt valitud olulisusnivoo määrab jaotuses ära kriitilise väärtuse, millest suuremaid χ^2-väärtusi võib pidada antud olulisusnivool statistiliselt olulisteks (sest tõenäosus neid juhuslikult saada on väiksem kui 5% või 1% või ...). Kriitiliste väärtuste leidmiseks tuleks kasutada vastavaid tabeleid.

Tabelist saame leida, et olulisusnivool 5% ning vabadusastme 1 puhul on χ^2-u kriitiliseks väärtuseks 3,84. Võrreldes näites saadud väärtust (4,17) kriitilise väärtusega ilmneb, et saadud väärtus on suurem ning seetõttu võime väita, et tõenäosus nullhüpoteesi kehtimiseks on alla 5% (vt joonis). See lubab meil nullhüpoteesi ümber lükata ning lugeda olulisusnivool 5% tõestatuks alternatiivse hüpoteesi, mille võiks sõnastada järgmiselt: Tallinna meessoost immigrandid on kohalikust poliitikast enam huvitatud kui naissoost immigrandid (või lihtsalt, et kohalikust poliitikast huvitatud nais- ja meessoost immigrantide osakaal on erinev).

Kui kasutate χ^2-testi tegemisel arvuti abi, siis pääsete tabelitest kriitilise väärtuse otsimisest, kuna arvuti suudab kiirest arvutada iga konkreetse χ^2-u olulisustõenäosuse, mis ütleb meile täpselt, kui suure tõenäosusega võiks meie erinevus olla tekkinud nullhüpoteesi olukorrast juhuse tõttu. Seda olulisustõenäosuse näitajat (*Significance*) tulekski nüüd võrrelda olulisusnivooga ning juhul kui ta on viimasest väiksem, siis võime nullhüpoteesi ümber lükata (st. oleme näidanud, et nullhüpoteesi kehtimise tõenäosus on väga väike ning seetõttu võime arvata, et tegelikult kehtib nullhüpoteesile vastupidine olukord).

Võib-olla panite tähele, et χ^2-test ei arvesta valimi suurust. Seepärast on tema kasutamisel mõned eeltingimused:
1. Objektide arv ei tohi olla alla 40ne
2. Ühegi lahtri oodatav sagedus ei tohi olla väiksem kui 1
3. Oodatav sagedus ei tohi olla väiksem kui 5 üle 20% lahtritest

Kui need tingimused ei ole täidetud või kui vabadusastmete arv on 1, siis **soovitatakse** kasutada χ^2-u arvutamisel Yaets'i parandust, mis seisneb erinevuse vähendamises.

$$\text{parandatud } \chi^2 = \sum \frac{\left(|T - O| - 0{,}5\right)^2}{O}$$

Meie näites oleks parandatud χ^2-u väärtuseks:

$$\chi^2 = \frac{\left(5 - 0{,}5\right)^2}{30} + \frac{\left(5 - 0{,}5\right)^2}{90} + \frac{\left(5 - 0{,}5\right)^2}{20} + \frac{\left(5 - 0{,}5\right)^2}{60} \approx 2{,}25$$

mis on väiksem kui kriitiline väärtus olulisusnivool 5% (3,84) ning seetõttu tuleks nüüd jääda nullhüpoteesi juurde ning tõdeda, et olulisusnivool 5% ei õnnestunud tõestada, et Tallinna meessoost immigrandid on kohalikust poliitikast enam huvitatud kui naissoost immigrandid.

8. Nähtustevahelised seosed.

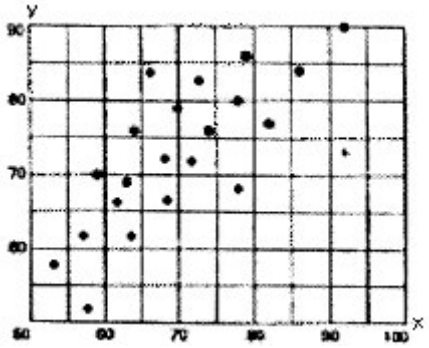
8.1 Korrelatsioon

Igapäevasest elust on seose mõiste meile hästi tuttav ning intuitiivselt tunnetatav. Arvan, et kõik saavad aru, mida ma mõtlen öeldes, et lapse vaimne areng on seotud tema vanusega, kuid ka sotsiaalsete tingimustega milles ta kasvab, sellega kui palju temaga tegeldakse või milline on tema IQ. Samuti on omavahel seotud näiteks ilmastikutingimused ja viljasaak või õppimisele kulutatud aeg ning eksamitulemus (võib-olla mitte alati, kuid üldreeglina siiski).

Samuti peaks olema selge, et seose tugevus võib olla erinev: võib arvata, et keskmine hinne lõputunnistusel on väga tihedalt seotud õpilase üldise intelligentsuse ja võimekusega, samas võib ta aga mingil määral olla seotud ka õpilase hoolsusega, tema tervisliku seisundi või koduste tingimustega.

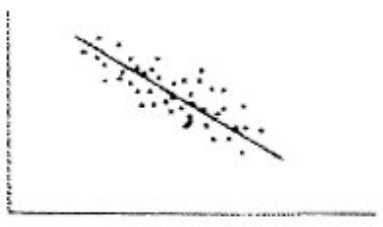
Seoste lähemaks uurimiseks pakuvad võimalusi korrelatsioon- ja regressioon- analüüs. Selleks, et kahe tunnuse vahel valitsevast seosest paremat ülevaadet saada, tuleks koostada korrelatsiooniväli, mis kujutab endast kahemõõtmelist punktdiagrammi, kus uuritavad kaks tunnust määravad ära teljed ning igat katseisikut (objekti) tähistatakse ühe punktikesega.

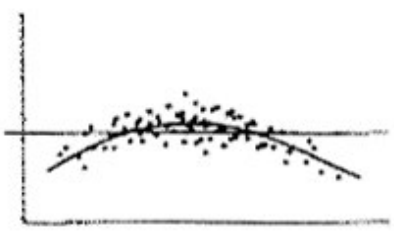
Vaatleme näiteks seost teoreetilisi teadmisi kontrolliva testi ja praktilisi oskusi kontrolliva testi vahel:

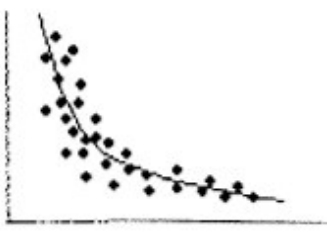


Kahe nähtuse vahel esineva seose iseloomustamiseks peame pöörama tähelepanu kolmele erinevale aspektile: **seose kujule, seose tugevusele ja seose suunale**.

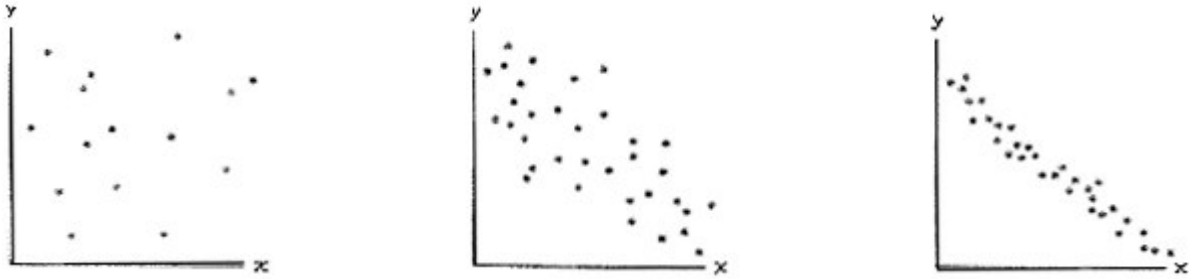
Seose kuju kahe nähtuse vahel määrab geomeetriline joon, millele punktide parv kõige lähedasem on. Kõige sagedamini on selleks sirgjoon ning sel puhul räägitakse **lineaarsest seosest**. Kuid võib ette tulla ka teistsuguse kujuga seoseid, mille puhul punktiparve iseloomustamiseks sobib paremini mingi kõverjoon. Vaata allolevaid jooniseid ning proovi leida iga joonise jaoks sobivad tunnused (st sellised tunnused, mille seos võiks vastata enam-vähem toodud kujule).







Seose tugevusest oli juba natuke juttu, kuid kuidas võiks see väljenduda meie korrelatsiooniväljal?
***
Tõepoolest, mida tugevam seos, seda tihedam (joone lähedale koonduvam) on punktiparv. Olukorda, kus kõik punktid koonduvad ühele joonele (st seos on täielik) nimetatakse funktsionaalseks seoseks. Selliseid seoseid te vaevalt eluliste nähtuste vahel leiate, kuid üks (elu)valdkond tegeleb siisk põhiliselt just selliste seostega ja see on loomulikult matemaatika. On ju kõigile teada, et (igale) ringi raadiusele vastab täpselt üks kindel ringi pindala jne.
Mittetäielikke seoseid nimetatakse vastavalt korrelatiivseteks seosteks.



Seose erinevaid suundi iseloomustavad järgmised kaks joonist:

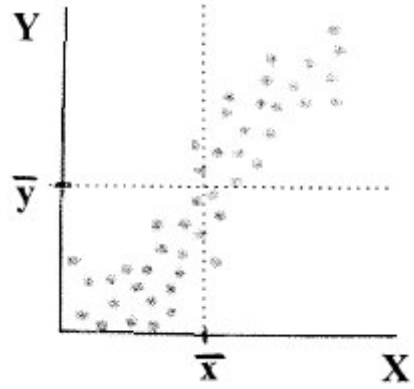
Kas oskate nende jooniste põhjal öelda, milles seose suund väljendub.
***
Seose suund loetakse positiivseks, kui ühe tunnuse väärtuste kasvades kasvavad ka teise tunnuse väärtused ning negatiivseks, kui ühe tunnuse väärtuste kasvades teise tunnuse väärtused kahanevad.

Lineaarset seost iseloomustavaks arvnäitajaks on **Pearson'i korrelatsioonikordaja r**, mis saadakse järgmise valemi abil:

iga katseisiku tulemuse erinevus keskmisest tõlgitakse universaalsele skaalale võttes ühikuks standardhälbe

$$r = \frac{\sum_{i=1}^{N} \frac{x_i - \bar{x}}{st.h._x} * \frac{y_i - \bar{y}}{st.h._y}}{N}$$

Antud valemit võib esitada ka mitmel erineval teisendatud kujul, kuid sellest kujust peaks olema kõige paremini näha korrelatsioonikordaja iseloom. Kõik põhineb iga üksiku katseisiku tulemuste võrdlemisel keskmiste tulemustega. Kui katseisik on mõlema tunnuse osas ülal- või allpool keskmist, siis saadakse korrutise väärtuseks positiivne arv, kui ta on aga ühe tunnuse osas ülalpool ning teise tunnuse osas allpool keskmist, siis saadakse korrutiseks negatiivne arv (vt joonist!).



Märk $\sum$ ütleb, et kõigi katseisikute jaoks saadud korrutised tuleb lõpuks kokku liita. Näeme, et kui I ja III veerandis on punkte rohkem kui II ja IV veerandis, siis peaks korrelatsioonikordaja tulema positiivne arv, vastupidisel juhul aga negatiivne. Seose puudumisel on kõigis veerandites punkte enam-vähem ühepalju ning seetõttu läheneb korrelatsioonikordaja väärtus arvule 0 (positiivseid ja negatiivseid korrutisi on ühepalju). Tänu sellele, et erinevused keskmisest standardiseeritakse (jagatakse läbi standardhälbega) on teada ka korrelatsioonikordaja maksimaalne ja minimaalne väärtus: juhul, kui on tegemist täieliku positiivse seosega on korrelatsioonikordaja väärtuseks +1, juhul kui on tegemist täieliku negatiivse seosega on korrelatsioonikordaja väärtuseks -1. Niisiis:

r = 0	tähendab seose puudumist,
r = +1	tähendab täielikku positiivset seost ja
r = -1	tähendab täielikku negatiivset seost

Korrelatsioonikordaja tugevuse tõlgendamiseks on erinevad autorid pakkunud erinevaid määranguid. Vast kõige enam kasutatav ning ka kõige lihtsam liigitus oleks järgmine (J.Tähtinen, 1993):

r < 0.30	olematu, väga nõrk
r < 0.70	keskmise tugevusega
r > 0.70	tugev

NB! Toodud on korrelatsioonikordaja absoluutväärtuste tähendused, sest korrelatsioonikordaja märk tugevust ei väljenda (vt eespool)!
D.Rowntree (1981,1991) on pakkunud välja täpsema jaotuse:

0.0 - 0.2	olematu, väga nõrk
0.2 - 0.4	nõrk
0.4 - 0.7	keskmine
0.7 - 0.9	tugev
0.9 - 1.0	väga tugev

Selleks, et korrelatsioonikordaja väärtust paremini mõista ja tõlgendada, tuleks teada, et korrelatsioonikordaja ruudul on omaette tähendus ning teda nimetatakse determinatsioonikordajaks. Determinatsioonikordajat kasutatakse tavaliselt olukorras, kus üks tunnus on teisest sõltuv. Determinatsioonikordaja näitab missugune osa sõltuva muutuja Y varieerumisest (e muutumisest) on seotud sõltumatu muutuja X varieerumisega. Determinatsioonikordajat tähistatakse tähega d:

$$\mathbf{d = r^2}$$

Kõige selgema ettekujutuse saame determinatsioonikordajast siis, kui me teda protsendina käsitleme (st korrutame saja protsendiga).

Oletame näiteks, et katsetulemuste põhjal saime lapse ja isa haridustasemete vaheliseks seoseks r = 0,5. Siit d = 0,25 ehk 25%. Kuidas seda siis tõlgendada?
***
Võib öelda, et isa haridustase määrab 25% sellest, millise haridustaseme laps saavutab. Ülejäänud 75% lapse haridustaseme muutumisest kirjeldavad aga mingid muud tunnused.

Tuleb silmas pidada, et korrelatiivne seos kahe nähtuse vahel ei tähenda ilmtingimata põhjuslikku seost, kuigi ta võib selle võimalikkusele viidata, eriti siis, kui seos on tugev. Seega, korrelatsiooni abil ei saa iialgi tõestada põhjusliku seose olemasolu kahe nähtuse vahel. Seda, et üks nähtus on teise põhjustajaks saab tõestada vaid põhjaliku kvalitatiivse, s.o uuritava kahe nähtuse sisulise analüüsi teel.
Asi on nimelt selles, et tegelikkuse nähtused on enamasti kompleksnähtused, kus kahe antud nähtuse vahelise seose määrajaks on sageli (ka) teised tegurid peale vaadeldava kahe. Seega, vaadeldavad kaks tunnust võivad mõlemad olla mõjutatud ühest kolmandast tunnusest.

Edasi tuleb meeles pidada, et kahe nähtuse vaheline korrelatiivne seos, ükskõik kui tugev see seos poleks, ei anna meile mingit informatsiooni suuruste absoluutväärtuste kohta st arvutades kummagi tunnuse aritmeetilise keskmise võivad need olla väga erinevad. Korrelatsiooni suureks vooruseks ongi see, et seosekordaja võib arvutada täiesti erinevatel skaaladel mõõdetud tunnuste vahel. Pikkus ja kaal on vaieldamatult omavahel seotud, kuigi ühte neist mõõdetakse sentimeetrites, teist kilogrammides jne.

Lõpetuseks tahan märkida, et Pearson'i korrelatsioonikordajat r võib kasutada vaid siis, kui tegemist on kahe numbrilise tunnusega (sest seosekordaja arvutamisel kasutatakse keskmisi ja standardhälbeid). Järjestusskaalal mõõdetud tunnuste puhul tuleks kasutada teisi seosekordajaid, mille arvutamisel kasutatakse näiteks järjenumbreid (st järjestatakse katsisikud kummagi tunnuse alusel ning võrreldakse saadud järjestusi). Üks tuntumaid mitteparameetrilisi seosekordajaid on Spearman'i ρ (loe roo), mis arvutatakse järgmise valemi järgi:

$$\rho = 1 - \frac{6\sum_{i=1}^{N}\left(jnr(X)_i - jnr(Y)_i\right)^2}{N^3 - N}$$

kus:
jnr(X) tähistab objekti järjekorranumbrit tunnuse X osas ning
jnr(Y) tähistab objekti järjekorranumbrit tunnuse Y osas.

Kui vähemalt üks tunnustest on nominaalne, siis tuleks kasutada χ^2- testiga sarnaseid meetodeid, milles lähtutakse risttabelist ning selle lahtrites olevate sageduste võrdlemisest.

8.2 Korrelatsioonikordaja statistiline olulisus

Nii nagu iga teisegi arvnäitaja korral peaksime enne korrelatsioonanalüüsi tulemuste üldistamist küsima, kui suur on viga, mille ma üldistades teen ning kas valimi andmete põhjal saadud korrelatsioon lubab mul mingi tõenäosusega väita, et selline korrelatsioon (st seos) ka üldkogumis olemas on.

Korrelatsioonikordaja standardvea saame arvutada järgmise valemi abil:

$$\text{st.viga}_r = \frac{1 - r^2}{\sqrt{N}}$$

Selleks, et hinnata üldkogumi korrelatsioonikordajat R tuleb aga leida korrelatsioonikordaja usaldusintervall. Korrelatsioonikordaja usaldusintervalli etteantud usaldusnivool saame eelnevast tuttaval viisil liites ja lahutades valimi korrelatsioonikordajale sobiva konstandiga korrutatud standardvea:

95% tõenäosusega kuulub R vahemikku	$\mathbf{r \pm 2*st.viga_r}$
99% tõenäosusega kuulub R vahemikku	$\mathbf{r \pm 2{,}5*\ st.viga_r}$

Seega, kui 100-liikmelise valimi korral saame korrelatsioonikordajaks r = 0,4 siis võime öelda, et 95% tõenäosusega kuulub üldkogumi korrelatsioonikordaja R vahemikku 0,4 ± 2*0,084 ehk vahemikku 0,232 kuni 0,568.

Vaatame nüüd kuidas on võimalik kindlaks teha, kas korrelatsioonikordaja on statistiliselt oluline või mitte (st kas valimis saadud korrelatsioon peegeldab tegelikku üldkogumis olevat seost või on ta nii väike, et võib olla tekkinud lihtsalt juhuse tõttu). Meie eesmärgiks on tõestada, et vaadeldavad kaks tunnust on üldkogumis seotud. Selleks on meil jälle vaja nullhüpoteesi:

H_0: R=0 ehk üldkogumis korrelatsioon puudub

ning olulisusnivood (mis näitab kuivõrd väike peab olema tõenäosus saada valimi korrelatsioonikordaja juhuse tõttu kui üldkogumis tegelikult korrelatsioon puudub selleks, et nullhüpotees ümber lükata)

Edasi peaks konstrueerima kõikvõimalike nullhüpoteesi olukorras juhuslikult tekkivate valimi-korrelatsioonikordajate jaotuse. Milliseid korrelatsioone võite oodata kõige enam?

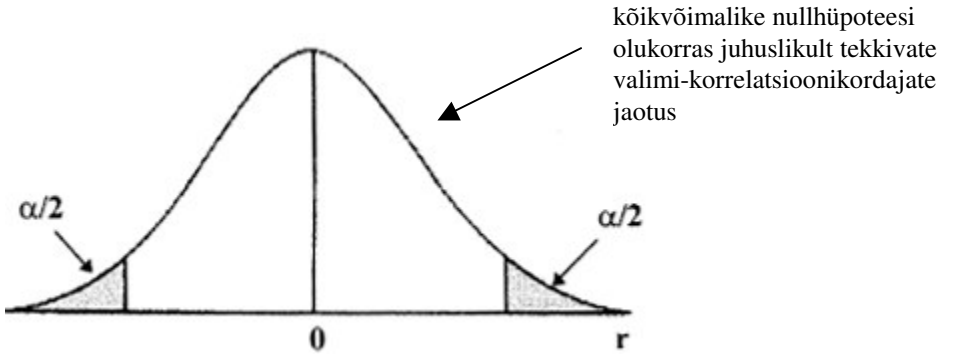
***

Kuna eelduseks on, et üldkogumis korrelatsioon puudub (R=0), siis juhuse tõttu saaksime kõige enam nullilähedasi valimi-korrelatsioone r.

Milline võiks olla saadava jaotuse kuju?

***

Kõigi ootuste kohaselt võiks sellel jaotusel olla normaaljaotusele lähedane kuju. Niisiis saame joonise:



Selleks, et võrrelda meie konkreetses valimis saadud korrelatsioonikordajat antud jaotusega, peaksime arvutama korrelatsioonikordaja standardvea väärtuse.
Tuletame meelde, et:

$$\text{st.viga}_r = \frac{1-r^2}{\sqrt{N}}$$

Kuna meil on praegu tegemist olukorraga, kus korrelatsioon puudub (R=0), siis saame standardvea valemist:

$$\text{st.viga}_r = \frac{1-0^2}{\sqrt{N}} = \frac{1}{\sqrt{N}}$$

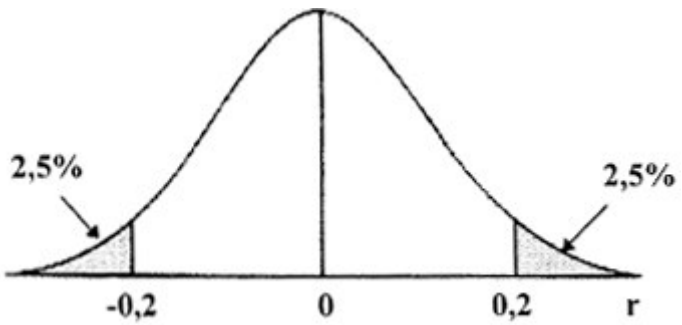
Näeme, et see kui suur on kriitilise korrelatsioonikordaja väärtus, sõltub ainult valimi suurusest. seega on lihtne leida kriitilise korrelatsioonikordaja väärtus mistahes suurusega valimi jaoks.

N. Kui N=36 ja olulisusnivoo on 5%, siis korrelatsioonikordaja kriitiline väärtus on r = $2*\frac{1}{\sqrt{36}} \approx 0{,}33$

Kui N=50 ja olulisusnivoo on 1%, siis korrelatsioonikordaja kriitiline väärtus on r = $2{,}5*\frac{1}{\sqrt{50}} \approx 0{,}36$

Kui N=100 ja olulisusnivoo on 5%, siis korrelatsioonikordaja kriitiline väärtus on r = $2*\frac{1}{\sqrt{100}} \approx 0{,}2$

Kanname viimase neist joonisele:



Näeme, et niipea kui valimi korrelatsioonikordaja absoluutväärtus on suurem 0,2'st, siis võime teda olulisusnivool 5% statistiliselt oluliseks pidada st võime 95% kindlusega väita, et mingisugune (valimiga samasuunaline) seos ka üldkogumis valitseb.

Loogilise arutluse põhjal näeme, et mida suurem valim, seda väiksem peab olema korrelatsioonikordaja, et ta oleks statistiliselt oluline. Näiteks, kui valimi suuruseks on 1000 objekti, siis kriitiline väärtus olulisusnivool 5%

on $r = 2^* \frac{1}{\sqrt{1000}} \approx 0{,}06$.

Võtame ühe näite:

Eesti elanikkonda esindavas 1000 liikmelises valimis läbi viidud küsitluse põhjal saadi, et inimeste haridustaseme ning õnnelikkuse vaheline korrelatsioon on r = 0,12.

Kuidas seda tulemust tõlgendada, kui meie eesmärgiks on teha järeldusi kogu Eesti elanikkonna kohta?

***

Osutub, et olulisusnivool 5% on see korrelatsioonikordaja statistiliselt oluline, seega võime väita, et meie valimi korrelatsioonikordaja peegeldab üldkogumis tegelikult esinevat seost õnne ja haridustaseme vahel.

Kas võime siis teha järelduse, et seos haridustaseme ja õnnelikkuse vahel on Eesti elanike hulgas tugev?

***

Loomulikult ei või! Seose tugevust iseloomustab ju korrelatsioonikordaja absoluutväärtus (mis meie valimis on 0,12). Eelnevast teame, et seoseid, mille tugevus on alla 0,3 tuleb lugeda väga nõrkadeks.

Selleks, et saada paremat ettekujutust üldkogumi korrelatsioonikordajast võiks arvutada veel korrelatsioonikordaja usaldusintervalli näiteks usaldusnivool 95%:

95% tõenäosusega kuulub R vahemikku $0{,}12 \pm 2^* \frac{1-0{,}12^2}{\sqrt{1000}} \approx 0{,}12 \pm 0{,}06$ ehk

95% tõenäosusega võime väita, et seose tugevus haridustaseme ja õnnelikkuse vahel Eesti elanikkonnas on vahemikus 0,06 kuni 0,18 (mis ei anna meile põhjust väita, et need kaks nähtust oleks omavahel tugevalt seotud).