**Katseandmete analüüs**

**Loeng 1. Sissejuhatavat ja kirjeldavat**

**Kirjutas Toomas Tammaru; konsultant Ants Kaasik, versioon sept 2025**

**Milleks statistikat (just eriti) bioloogias vaja on?** Aga sellepärast, et ka mingi selgeltki piiritletud rühma moodustavad **bioloogilised objektid pole kunagi täpselt ühesugused**. Nii pole kaks niidu-uruhiirt kunagi täpselt samasugused, nad on erinevad tänu oma isendiomase genotüübi ja läbitud ontogeneetilise arengu iseärasustele. Samamoodi pole täpselt ühesugused ka nende hiireorganismide osad (nt munasarjad) ja ammugi pole seda erinevad niidu-uruhiire populatsioonid. Häda on aga selles, et midagi bioloogliselt sisukat **väita tahame** me tavaliselt mitte konkreetse uruhiireisendi (munasarja, populatsiooni) kohta, vaid enamasti ikka **millegi üldisema**, näiteks kogu liigi **kohta**. Kogu liiki (st kõiki isendeid) me uurida ei saa, uurida saame vaid mingit piiratud hulka neist. Sellise kogu liigi kohta väitmise vaid piiratud hulga uuritud isendite põhjal teebki just raskeks isendite (vms objektide) individuaalne varieeruvus. Statistika ongi selleks välja mõeldud, et sellest individuaalsete eripärade pilvest läbi näha ja langetada põhjendatud otsus, mida me ikkagi selle kogu liigi kohta väita saame.

 Selles osas erineb bioloogia oluliselt füüsikast ja keemiast, kus suure hulga probleemide lahendamisel võib katseobjekte identseteks lugeda. Nii näiteks kui tahame uurida, kuidas reageerib üks mool väävelhapet kahe mooli naatriumhüdroksiidiga (nt palju soojust eraldub), siis pole põhjust arvata, et üks purgitäis puhast ainet reageerib teistmoodi kui teine - väävelhape on väävelhape ja purgitäite individuaalne omapära mängu ei tule. Siin võib põhimõtteliselt teha vaid ühe katse ja selle õnnestumise korral väita, et asi nii ongi. Teistsugune on olukord, kui tahame uurida, kas hiirepojad kasvavad paremini porgandit või kapsast süües. Siin sellest ei piisa, et anname ühele hiirele ühte ja teisele teist ja vaatame, kumb paremini kasvab - see, kes paremini kasvas, võis seda teha hoopis muul põhjusel kui toidu tõttu, seda siis tänu just sellele eespool seletatud bioloogilise objekti individuaalsele eripärale. Ühe hiirepoja kiirem kasv võis olla tingitud geneetilistest erinevustest nende poegade vahel või ka näiteks varem kogetust - et kui üks poeg näiteks oli näljas nooruses.

Intuitiivselt on selge, et kindluse saavutamiseks ses osas, et kuidas need hiirepojad siis üldiselt kasvavad kahe erineva aedvilja peal, tuleb katseid teha võimalikult paljude hiireisenditega, just nimelt selleks, et individuaalsete erinevuste tagant seda “üldist suunda” ära näha.

(Matemaatilise) statistika ülesanne ongi see intuitsioon rangeks muuta, ehk siis öelda, kui palju neid hiirepoegi on vaja uurida **ja kui kindlalt me miski arvu poegade puhul midagi väita saame**. Järgmises lõigus formuleerime sama mõtte rangemalt.

 Nii, **valimi** (***sample***) moodustavad need hiired, keda me reaalselt katsesse võtsime ja seal uurisime-mõõtsime. Valimi suurust (objektide arvu selles) tähistatakse üldiselt n tähega (või ka N). **Üldkogumi** moodustavad kõik need hiired, kellele me oma kaste tulemusi üldistada tahame, näiteks kogu maailma (või Saaremaa, või meie koduniidu) niidu-uruhiired. Järeldada tahame me midagi üldkogumi kohta, aga uurida saame vaid väiksest osa sellest - valimit. Statistika ongi selleks, et vastata küsimusele, mida me oma **valimi põhjal üldkogumi kohta järeldada** saame ja mida ei saa.

Alustame **kirjeldamisest**. Kõigepealt tuleb iseloomustada valimit. Edasi tuleb mõelda, mida kirjeldavat saab valimi põhjal üldkogumi kohta järeldada.

Valimis on hulk **objekte**, no näiteks siis hiiri. Iga objekti puhul on mõõdetud mingi muutuja (tunnuse) väärtused. Muutujad võib jagada pidevateks ja diskreetseteks. **Pidev muutuja** (*continuous variable*) võib saada misiganes arvulisi väärtusi (vähemalt teatud vahemikus), pidevad muutujad on näiteks pikkus ja kaal. Diskreetne (ehk kategooriline) muutuja võib saada vaid teatud piiratud arvu konkreetseid väärtusi, nii võib muutuja “sugu” saada väärtusi “emane” ja “isane”. Mingis valimis mõõdetud muutujate väärtused (ehk **vaatlused,** *observations*) moodustavad **jaotuse (***distribution***)**.

 Diskreetsete muutujate jaotusi on lihtne iseloomustada: loeme ära, palju nt emaseid ja isaseid on ja kogu lugu. Kui võimalikke väärtusi on kaks nagu selles näites, on tegu kaheväärtuselise (*binary*) jaotusega, kuid ette tuleb ka selliseid jaotusi, kus võimalikke väärtusi on rohkem. Nii võivad miski taime õied olla punased, valged või sinised. Loeme ära ja ongi iseloomustatud, milline oli tunnuse jaotus kõnealuses valimis.

 Pidevate tunnustega on keerulisem, kuid huvitavam. Kõige lihtsam viis jaotust graafiliselt iseloomustada oleks kanda väärtused punktidena arvteljele. Siiski pole selline viis eriti illustratiivne ja selgema pildi saamiseks jaotatakse väärtused vahemikeks ja vaatluste arvu miskis vahemikus kujutatakse tulba kõrgusena, tekib **histogramm** (joonis 1-1).

 Histogrammi abstraktsioon on tihedusfunktsioon. Sellise joonele läheneme, kui (kasvõi mõtteliselt) üha suurendame vaatluste hulka ja üha vähendame histogrammi tulpasid defineerivate lõikude pikkusi (joonis 1-2). Tihedusfunktsiooni abil saab arvutada, kui suur osa vaatlusi jääb teatud väärtuste vahemiku. Jaotuse lähendamine pideva funktsiooniga on eriti kasulik statistika teoorias, sest just pideva funktsiooniga on mugav matemaatikas opereerida (ühte ja teist tõestada).

Tuntuim pidev jaotus on **normaaljaotus**, selle tihedusfunktsioon on nn kellakujuline (joonis 1-3). Siiski on asjalik tähele panna, et sugugi mitte kõik umbes sellised jaotused pole normaaljaotused, normaaljaotuse tihedusfunktsioon on ikkagi konkreetne matemaatiline funktsioon, millel on konkreetne võrrand; kvalitatiivselt sarnase kujuga võivad olla ka paljude teiste funktsioonide graafikud.

 Normaaljaotus tekib siis, kui tunnuse väärtust mõjutavad väga paljud juhuslikud tegurid ja neist igaühe mõju on väga väike, nii see eluslooduses tüüpiliselt ka on. Normaaljaotus on teoreetiline abstraktsioon - eluslooduses ei ole ükski asi täpselt normaaljaotusega juba kasvõi sellepärast, et normaaljaotuse sabad lähevad lõpmatusse: kui usuksime, et karude kehakaal on rangelt normaaljaotusega, peaks olema võimalik, et looduses võib mingi tõenäosusega esineda ka negatiivse kaaluga ja üle kümne tonni kaaluvaid karusid. Siiski on paljud tunnused looduses normaaljaotusele (väga) lähedase jaotusega, millest praktikas piisab (sellest hiljem, kus seda normaaljaotusega võrdlemist vaja läheb).

Edasi vaatleme **pideva jaotuse kirjeldamist**.

Jaotust iseloomustavad

 - jaotuse „keskkohta“ iseloomustavad näitajad (igasuguseid valimeid või seoseid iseloomustavaid arve nimetatakse statistikas statistikuteks);

 - hajuvust iseloomustavad statistikud;

 - jaotuse keerulisemaid omadusi kirjeldavad statistikud.

**Erinevad “keskmised”**

ehk siis jaotuse keskkohta võib defineerida mitmel erineval moel.

 **(Valimi)keskmine (***mean***,** üldkogumi puhul kasutatakse terminit **keskväärtus**) ei vaja ilmselt kommentaare: liidad kokku, jagad kui mitu oli, ehk siis tegu on vaatluste tavalise aritmeetilise keskmisega;

 **Mediaan** on uuritava muutuja väärtus, millest suuremaid ja väiksemaid väärtusi on võrdselt (joonis 1-4);

 **Mood** on väärtus, mis iseloomustab kõige suuremat hulka objekte (joonis 1-5).

Sümmeetrilise jaotuse (näiteks normaaljaotuse) korral langevad keskväärtus ja mediaan (sageli ka mood) ühte, ebasümmeetrilise korral ei pruugi.

**Näide** (joonis 1-6): usutavasti on X riigi elanike sissetulek ebasümmeetrilise jaotusega, pika sabaga paremale. Pane tähele, et x-lane saab keskmiselt (keskväärtus) rohkem palka kui keskmine x-lane (mediaan), tüüpiline x-lane võib aga veel vähem palka saada (mood).

Igal neist kolmest on **erinev sisuline tõlgendus** ja kõiki neid kolme saab ja tuleb kasutada vastavalt vajadusele (probleemi iseloomule). Tuleb mõelda, mida täpselt teada tahame ja milline neist statistikutest vastab küsimusele sisuliselt kõige paremini. Nii on vist rahva elukvaliteedi uuringutes mõistlik kasutada **mediaani** (st kuidas elab keskmine x-lane) - nii ei avalda üksikud pururikkad mingit mõju (samas kui aritmeetilisele keskmisele nad seda avaldavad). Samas, kui tahame teada, palju laekub omavalitsusele inimese kohta makse, huvitab meid **keskmine**. (Nt kehapikkuse) **mood** võiks meid huvitada näiteks siis, kui ostame parkidesse pinke - tahame teada, kui kõrged nad olema peaksid, et võimalikult paljudel seal hea istuda oleks. Tõlgendused X-riigi näite varal veelkord: aritmeetiline keskmine näitab, milline on x-lase sissetulek keskmiselt; mediaan näitab, milline on keskmise x-lase sissetulek; mood näitab, milline on tüüpilise x-lase sissetulek – need kolm ei ole asja ei ole samad asjad!

 Valimi keskmine, mediaan ja mood on ka üldkogumi vastavate statistikute parimateks võimalikeks hinnanguteks, mida valimi põhjal teha saab. See tähendab seda, et kui valimi keskmine on 2,7, siis parim hinnang üldkogumi keskmisele on samuti 2,7 - st pole põhjust arvata, et valimi keskmine hindaks üldkogumi keskmist süstemaatiliselt üle või alla. See jutt võib praegu triviaalne ja mõttetu tunduda, kuid kohe näeme, et nt dispersiooni puhul see nii pole: valimi dispersioon ei ole üldkogumi dispersiooni parimaks hinnanguks.

Peale keskmise huvitab meid ka see, kuidas on üksikud vaatlused ümber keskmise jaotunud, ehk siis

**hajuvusstatistikud**

iseloomustavad seda, kui laia tunnuse väärtuste vahemikku on uuritava valimi objektid ‘laiali hajutatud’ (joonis 1-7). Hajuvust võib mõõta (matemaatiliselt väljendada) mitut moodi (nii nagu keskkohtagi sai mitut moodi defineerida), seepst siis erinevad statistikud, igaühel omad head ja vead.

Keskne hajuvusstatistik on **dispersioon** (***variance***). Valimi dispersioon on defineeritud valemiga



kus xi on üksikvaatlus (neid on N tükki, i on iga vaatluse number), müü on valimi keskmine.

*Dispersiooni võib arvutada ka nii, et üksikvaatluste ruutude keskmisest lahutatakse ükikvaatluste keskmise ruut, tulemus on sama.*

Siiski, valimi põhjal arvutatav hinnang kogu populatsiooni tegelikule dispersioonile on veidi teine.



Pane siis tähele, et dispersioon valimis hindab üldkogumi dispersiooni keskmiselt natuke alla (ehk siis ei ole parim hinnang), aga vastav parandus on lihtne sisse viia ja et erinevus muutub tühiseks suurte valimite korral. Aga igal juhul oleme siin jõudnud selleni, mis on statistika tuumaks - valimi põhjal üldkogumi kohta järelduste tegemiseni.

Siin oluline kõrvalepõige: võimaldamaks valimi põhjal üldkogumi põhjal järeldusi teha, peab **valim olema võetud kogu üldkogumist juhuslikult**. Ehk siis kui tahame midagi väita kogu Eesti niidu-uruhiirte kohta, tuleb meid hiiri valimisse püüda juhuslikult üle Eesti. Kui püüame hiiri ainult Saaremaalt, saame selle saaremaise valimi põhjal teha järeldusi ikka ainult Saaremaa niidu-uruhiirte kohta.

 Teema juurde tagasi ja pane tähele, et dispersiooni ühik pole sama, mis muutuja enda ühik: vt valemit, x läheb seal ruutu! Nt kui kala pikkust mõõdame meetrites, siis selle muutuja dispersiooni mõõdame ruutmeetrites. Seega ei saa dispersiooni kasutada pluss-miinus tüüpi väljendites (kaks pluss-miinus üks) hajuvuse hinnanguna, sest liita ja lahutada saab ju vaid sama ühikuga suurusi. Samuti ei saa me joonisele dispersiooni joonistada (või kui, siis oleks see kummaline koledasti).

Miks siis dispersiooni ikkagi kasutatakse? No eks sellepärast, et tal on muid meeldivaid omadusi statistiliste operatsioonide tarbeks, üks hea omadus on kindlasti aditiivsus, st **dispersiooni saab komponentideks lahutada** ja kokku annavad need komponendid (*teatud tingimustel (muutujate mittekorreleeritus), tõsi küll, millel siin lähemalt ei peatu*) kogu dispersiooni kokku. Kui kasutame hajuvuse mõõduna dispersiooni, saame öelda, et niimitu protsenti hajuvusest läheb selle asja arvele ja naamitu teise asja arvele, nt saame asja uurida ja öelda, et kartulisaagi varieeruvusest talude vahel läheb 20% väetamise erinevuste arvele, 20% ilma arvele ja 60% talumehe töökuse arvele. Pane tähele, et mitte niimitu protsenti muutuja väärtusest (kartulisaagist), vaid selle dispersioonist. Dispersiooni komponentideks jagamisel põhineb kogu dispersioonanalüüs (vt edaspidi) ja näiteks ka kvantitatiivne geneetika (asjast räägitakse evolutsioonilise ökoloogia loengus).

**Näide**. Kala pikkus on pea pikkuse, keha pikkuse ja saba pikkuse summa. Näeme, et pea, keha ja saba pikkuse väärtuste dispersioonid annavad kokku kala pikkuse dispersiooni, just nii nagu vastavad väärtused ise seda teevad. Standardhälvetel ja muudel värkidel sellist aditiivsuse omadust ei ole. Dispersiooni komponente on mõistlik protsentides väljendada.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| kala | pea | keha | saba | kokku |
| 1 | 3 | 8 | 4 | 15 |
| 2 | 4 | 7 | 4 | 15 |
| 3 | 5 | 10 | 4 | 19 |
| 4 | 6 | 7 | 4 | 17 |
| 5 | 7 | 8 | 4 | 19 |
| disp | 2 | 1,2 | 0 | 3,2 |
| % | 62,5% | 37,5% | 0 | 100% |

**Standardhälve (*standard deviation*, SD) on ruutjuur dispersioonist.** Tema eelis on seega see, et teda mõõdetakse samades ühikutes kui muutujat ennast, seega saab teda tunnuse väärtustele liita (pluss-miinus) ja joonisele panna. Kui kala pikkuse dispersiooni mõõdame ruutmeetrites, siis kala pikkuse standardhälvet meetrites nagu kala pikkust ennastki. Pane tähele, et kui standardhälve suureneb n korda, siis dispersioon suureneb n-ruudus korda, ehk siis kui SD suureneb kolm korda, siis dispersioon suureneb 9 korda. Seega peab täpsustama, mida mõeldakse, kui miski arv kordi suuremast hajuvusest räägitakse.

Kuidas standardhälvet mõista ehk näitlikustada, et kuipalju muutuja väärtus siis ikka varieerub kui standardhälve on nii või naa suur? Aga nii, et kui on tegu normaaljaotusega, siis vahemikku keskmine miinus üks standardhälve kuni keskmine pluss üks standardhälve jääb 68% vaatlusi (joonis 1-8), seega pea kolmandik vaatlusi on keskmisest kaugemal kui üks standardhälve. Seega, kui taimemäärajas on kirjas, et päevalille kõrgus on 3 pluss-miinus 1 ja esitatud hajuvusstatistik on standardhälve (loodetavasti on määraja autor seda selgelt öelnud!), siis leitud 4,1 m kõrguse päevalillega kuulsaks ei saa, see ei ole kindlasti rekord: 16% päevalilledest on kõrgemad kui 4 meetrit.

 **Variatsioonikoefitsient** (***coefficient of variation,* CV**) on standardhälve jagatud tunnuse keskmisega. See võimaldab võrrelda erinevate tunnuste suhtelist varieeruvust, sobiv väljendada protsentides. Nii näiteks võib öelda, et lämmastikusisaldus varieerus kartulimugulate vahel rohkem (CV=26%) kui kaal (CV=17%), CV-s on seega võrreldavad ka siis, kui võrreldavad muutujad on väljendatud erinevates ühikutes.

 **Kvantiilid** ehk **fraktiilid** on muutuja väärtused, millest ühele ja teisele poole jääb (st on suuremad või vastavalt väiksemad) kindel % väärtusi meie valimis. Mediaan on 50% kvantiil. 25% kvantiilist väiksemad on 25% vaatlusi valimis. 25% ja 75% kvantiile nimetatakse **kvartiilideks** (joonis 1-9). Kvantiilide oluline eelis on see, et nende abil saab iseloomustada ebasümmeetrilisi jaotusi ja jaotuse ebasümmeetrilisust. Seda siis sellepärast, et keskmisest (või mediaanist) 75% kvartiilini ei pruugi sugugi olla samapalju maad kui 25% kvartiilini, seega meil on kaks erinevat arvu ja nende (st kvartiilide kaugus mediaanist) erinevus näitab ebasümmeetrilisust (kui iseloomustame hajuvust pluss-miinus SD-ga, siis iseloomustame seda ühe arvuga (SD), mis siis „kehtib mõlemas suunas“). Eespool seletatust järeldub, et väärtused keskmine miinus SD ja keskmine pluss SD on kvantiilidena tõlgendatavad normaaljaotuse korral (sellesse vahemikku jääb 68% vaatlustest), üldjuhul ei ole, st üldjuhul ei ole otsest seost kvantiilide ja standardhälbe vahel.

 Pane hästi tähele, et kõigi ülaltoodud hajuvusstatistikute **väärtused ei sõltu süstemaatiliselt valimi suurusest**, st mida suurem valim, seda täpsemad on mõistagi hinnangud, kuid mingit trendi (suuremaks või väiksemaks) valimi suuruse muutusega ei kaasne (joonis 1-10, vrd kohe allpool kus see nii ei ole).

**Hajuvusstatistikutele** **sarnased**, kuid **sisuliselt väga erinevad** näitajad iseloomustavad meie **teadmiste täpsust** üldkogumi keskmisest. Nad pole enam mõeldud kirjeldama üldkogumis esinevat hajuvust, kuigi nad mõistagi sõltuvad sellest.

**Standardviga** **(*standard error,* SE)** ehk valimi keskmise standardhälve on SD/√n (jagatud ruutjuur-n-ga). *Formaalselt on tegu standardhälbega sellises uues üldkogumis, mis tekib, kui tegelikust üldkogumist võetakse uuritava valimiga võrdse suurusega valimeid ja arvutatakse uute valimite keskmised. Standardviga on siis selliste hüpoteetiliste valimite keskmiste standardhälve – see jutt sellele, keda formaalne definitsioon huvitab, eksamil ei küsi*. See selleks, aga ta iseloomustab siis **meie teadmiste täpsust** uuritava üldkogumi keskmisest, mida täpsem on meie teadmine, seda väiksem on SE. SE sõltub seega a) üldkogumi dispersioonist; b) valimi suurusest. Mida suurem on valim, seda väiksem on SE. Valimi suurenedes läheneb SE nullile (joonis 1-10). See on siis oluline erinevus SD-st.

 **Usaldusintervall** (***confidence interval of the mean***) on SE-ga analoogiline kuid otsesema tõlgendusega (aga keerulisemalt arvutatav) suurus. Definitsioon on lihtne: kui tegelik keskmine asuks väljaspool usaldusintervalli, oleks antud keskmisega valimi saamine väiksema tõenäosusega miskist ette antud väärtusest. Tavaliselt kasutatakse 95% usaldusintervalle, ehk siis kui joonistame valimi keskmise ümber usaldusintervalli, siis märgime ära väärtuste vahemiku, milles üldkogumi keskmise asumine on igati ootuspärane (NB statistika põhituum – valimi põhjal üldkogumi kohta järelduste tegemine; joonis 1-11). Veidi üle nurga väljendudes võib öelda, et üldkogumi tegelik keskväärtus võiks tõenäoliselt asuda usaldusintervalli sees, sellest väljaspool ta tõenäoliselt ei asu. Suurte valimite korral läheneb vahemik mean pluss-miinus SE (st keskmisest üks standardviga kummalegi poole) 68% usaldusintervallile, seega on see standardveaga defineeritud vahemik oluliselt kitsam kui 95%-line usaldusintervall. Usaldusintervall on sümmeetriline valimi keskmise suhtes ja muidugi on ta seda kitsam, mida suurem on valim – suurema valimi korral teame üldkogumist kindlasti rohkem kui väiksema korral.

 Moraal veelkord: SE ja usaldusintervallid iseloomustavad meie teadmist populatsiooni keskmisest, nad ei ole mõeldud kirjeldama hajuvust üldkogumis.

**Hajuvusstatikute esitamine**

 Teaduskirjanduses on üldlevinud ja enamasti nõutav, et keskmised (ja muud valimi parameetrid või üldkogumi parameetrite hinnangud) esitatakse koos miskite hajuvusstatistikutega, tekstis tehakse seda siis pluss-miinus märgi abil enamasti, joonisel *error bar*’ide kujul. Eelnenust peaks selge olema, et alati **tuleb ära mainida** (näiteks ja eriti tabeli peas ja joonise tekstis), **millise hajuvusstatistikuga on tegu** (enamasti kas SD või SE), see ei ole iseenesest mõistetav.

 SD’d kasutame siis, kui eesmärgiks on kirjeldada varieeruvust, SE’d siis, kui esituse eesmärk on näidata ja võrrelda keskmisi. Tulpdiagrammil (joonis 1-12) SE’sid vaadates saame kiire esialgse ülevaate erinevuste ligikaudsest statistiliselt olulisusest (vt järgmine loeng) - kui umbes 1,3-ga korrutatud SE’de otsad ei kattu, on erinevus oluline, kui kattuvad, siis pole. Loomulikult ei asenda selline vaatamine asjakohaseid teste (vt kohe järgmine loeng), aga esmapilguülevaate jaoks hea küll.

 Tugevalt **ebasümmeetrilise** jaotuse puhul tuleb eelistada sümmeetrilistele hajuvusstatistikutele näit. **kvantiile**, seda eriti sellisel puhul kui pluss-miinus SD läheb ühest otsast üle sisuliselt asjaliku piiri. Nii võib ebasümmeetrilise jaotuse puhul kergesti juhtuda: näiteks kui uurime parasitismi protsenti eri putukapopulatsioonides, siis olukorras 0,2 0,1 0,1 0,1 0,9 0,8 saame 0.36 pluss-miinus 0.37 (SD), seega see miinusega pool kipub alla nulli, mis pole antud kontekstis mõistlik väärtus. Kvantiilidega sobib kokku mediaan, mitte aritmeetiline keskmine; on ju mediaan ise ka kvantiil, 50% kvantiil täpsemalt.

 ***Box plot*** (karpdiagramm, joonis 1-13) on sobiv lahendus juhuks, kus soovitakse esitada mitut hajuvusstatistikut korraga. Karbi ülemine ja alumine piir võivad siis näiteks näidata SE’d, vurrud SD’d. Sümmeetrilisel juhul pole keskmist kriitiline joonisele panna, on see ju siis täpselt keskel. Ebasümmeetrilise jaotuse puhul peaks kasutama kvantiile (näiteks 25% ja 75% karbi piirid ja 90% vurrud). Siin pole mediaan enam mõistagi alati keskel ja tuleks ära märkida, näiteks joonega. Aga - igal juhul, **karpdiagrammi esitamisel tuleb kindlasti lahti seletada**, mis statistikuid miskid jooned tähistavad.

 Tavalised tulpdiagrammid on karpdiagrammidele eelistatavad juhul, kui on oluline näitlikustada võrreldavate rühmade keskmiste suhet (mitte vahet), st siis on oluline ära näidata keskmiste kaugus nulljoonest.

**Peale keskmiste ja hajuvuse on jaotusel mõistagi ka keerulisemaid omadusi.**

Jaotuse asümmeetrilisust mõõdab statistik asümmeetriakordaja (***skewness***). Sümmeetrilise jaotuse (nt normaaljaotuse) *skewness* on null. Kui pikk saba on paremale (suuremate väärtuste poole), on *skewness* positiivne. Vasakule sabaga siis vastavalt negatiivne.

Jaotuse järsakus (***kurtosis***) väljendab jaotuse tipu teravust. Normaaljaotuse järsakus on null. Terava tipuga jaotustel on *kurtosis* positiivne, lamedatel negatiivne (vt tahvlile).

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* jutu lõpp \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*