Peatükk 1

Andmetest

Siin peatükis näeme, kui keerulise kõlaga sõnu saab kasutada andmetest rääkimisel; vaatame, kuidas algandmeid üles kirjutada ja kuuleme, milliseid silte võib mõõtmistulemustele külge kleepida

ehk

Andmemaatriksist, tunnuste tüüpidest ja kasutatavast tähistusest, väärtuste kodeerimisest ja puuduvatest väärtustest.

Selleks, et saaksime midagi objektiivset väita meid huvitava nähtuse, protsessi või objekti kohta on vaja vaatlus- või katsetulemusi. Enamasti pakub meile rohkem huvi see, milliseid järeldusi ja üldistusi me nende mõõtmistulemuste põhjal teha saame, kui üks või teine konkreetne mõõtmistulemus ise. Sestap on kerge unustada algmaterjal ja pürgida kohe kõrgustesse ja keerukate analüüside risträgastikku. Paraku ei seisa ükski maja kindlalt, kui vundament on hooletult ehitatud. Käesolevas peatükis vaatleme, millist terminoloogiat saab kasutada algandmetest rääkides ja sedagi, kuidas algandmeid nii kirja panna, et algandmetel tuginevad analüüsid hiljem ka kriitikatuultes püsima jääksid.

1.1 Objekt ja tunnus

Paljud statistikaga seotud arusaamatused on välditavad, kui inimesed saaksid täpselt aru, mida või keda nad uurivad (või keda on uuritud artiklis, mida parasjagu loetakse).

Definitsioon 1.1 ***Objekt*** *on uurimisalune ühik, üksikindiviid.*

Objektideks võivad näiteks olla linnupojad või pesakonnad; puud, metsatukad või punktid metsas; põdrad või ristamiskatse tagajärjel sündivad olevused.

Vahel on ka samade andmete puhul olemas mitu erinevat võimalust valida uurimisobjekti. Näiteks vaatame situatsiooni, kus on vaadeldud kahes pesas koorunud linnulapsi — ühes pesas koorus 9, teises 1 linnulast. Valides uurimisobjektiks pesakonna (kurna), saame tunnuse "*pesakonna suurus*" keskmiseks 5 (keskmine pesakonna suurus on 5); valides uurimisobjektiks aga linnulapse, saame tunnuse "*pesakonna suurus*" keskmiseks 8,2 (linnulaste keskmine päritolupesakonna suurus on 8,2). Vaata ka tabelit 1.1.

Tabel 1.1: Samad andmed - aga uurimisobjekt on erinev

Objekt - linnulaps

Pesakonna nr	Pesakonna suurus
1	9
1	9
1	9
1	9
1	9
1	9
1	9
1	9
1	9
2	1

Keskmine pesakonna suurus: 8,2

Objekt - pesakond

Pesakonna nr	Pesakonna suurus
1	9
2	1

Keskmine pesakonna suurus: 5

Näiteks metsa uurides võivad uurimisobjektideks olla puud või punktid metsas (võrdle: "80% vaadeldud metsapuudest olid pajud" vs "35% vaadeldud metsa-aladest olid kaetud pajudega"); taimede produktiivsuse uurimisel võib objektideks valida näiteks kas taime päevased või nädalased juurdekasvud (päeva või nädala jooksul lisanduv biomass), kusjuures uuritava tunnuse stabiilsus võib märgatavalt sõltuda meie valikust (nädala jooksul lisanduvad biomassid on sarnased; ühe päeva jooksul lisanduv biomass aga on üsna varieeruv suurus); jne.

Definitsioon 1.2 ***Tunnus*** *on objekti iseloomustav näitaja, mida põhimõtteliselt on võimalik mõõta või vaadelda.*

Hiiri uurides võivad tunnusteks olla karvavärv, kaal, sabapikkus, liik ja vanus; taimi uurides võivad tunnusteks osutuda kasvukoht, pikkus, lehtede arv, biomass, liik jne.

1.1.1 Tähistustest

Tunnuste nimede kirjutamisel kasutame edaspidi suuri tähti, näiteks $VANUS$, $SABAPIKKUS$ või $LIIK$. Vahel võime kasutada ka lühendeid, näiteks tunnuse "hiire poolt aasta jooksul läbinäritud raamatulehekülgede arv" võime tähistada sümboliga X (pane tähele - kasutame ikkagi suurt tähte X). Konkreetsete mõõdetud väärtuste tähistamiseks kasutame aga väikeseid tähti. Näiteks tunnuse X väärtus ühe konkreetse hiire puhul on tähistatud sümboliga x. Kui soovime täpsustada, millisel konkreetsel objektil vastav mõõtmine on aset leidnud, kasutame objekti numbrit alaindeksis: x_3 on tunnuse X väärtus 3. objektil (näiteks 3. hiire poolt rikutud lehekülgede arv).

Suurte tähtedega võime tähistada ka tulevasi mõõtmistulemusi, mille väärtus arutelu hetkeks pole selgunud. Näiteks planeerides järgmisel aastal aset leidvat uuringut saab rääkida 3. hiire mõõtmistulemusest kui suurusest X_3 (mis võib osutuda milleks iganes), peale uuringu toimumist ja andmete kogumist aga juba kui mõõtmistulemusest x_3 (mis on üks konkreetne ja teadaolev number).

1.2 Andmemaatriks

Arvuti jaoks andmete mõistetavaks tegemisel tuleb algandmed sisestada arvutisse kindlal kujul. Enamik statistikaprogramme soovib, et andmed oleksid sisestatud nn. objekt-tunnus maatriksina, st. sellise tabelina, kus iga veerg kujutab ühte tunnust ja iga rida ühte objekti.

Näide 1.1 *Käidi kümnel põllul ja koguti andmeid mullatüübi, mulla niiskuse ja viljakuse kohta. Saadud andmemaatriks on esitatud tabelis 1.2.*

Kaks võimalikku objekt-tunnus maatriksit on esinenud juba ka tabelis 1.1.

Tähtis on meeles pidada, et ühe (uurimis)objekti kohta tohib objekt-tunnus maatriksis olla vaid üksainus rida.

Kui andmed pole statistikaprogrammi sisestatud objekt-tunnus maatriksina, siis võib karta, et varem või hiljem leiab aset inimlik eksimus ning

Tabel 1.2: Objekt-tunnus maatriks

Põld	Mullatüüp	niiskus	suvinisu viljakus (kg/ha)
1	savimuld	niiske	3624
2	liivsavimuld	paras	4782
3	savimuld	niiske	4274
4	liivmuld	kuiv	3927
5	savimuld	paras	4630
6	liivmuld	paras	4920
7	savimuld	niiske	4260
8	savimuld	paras	4935
9	liivsavimuld	paras	5035
10	liivmuld	kuiv	4500

analüüsi tegev inimene interpreteerib arvuti poolt väljastatavaid tulemusi valesti.

Märkus: vahel kasutatakse ka andmemaatrikseid, kus ühe objekti kohta on kirjas mitu rida (nn kordusmõõtmiseid sisaldavad andmestikud). Selliste andmemaatriksite analüüs nõuab eriliste statistiliste meetodite kasutamist (kordusmõõtmiste analüüs, *repeated measures analysis*, ...) ja isegi pealtnäha lihtsatele küsimustele vastamine (milline on keskmine?) võib õige vastuse leidmine osutuda vägagi keeruliseks ülesandeks. Sellisel kujul esitatud andmete analüüs nõuab suuri statistika-alaseid teadmiseid ja pole enamasti algajale jõukohane.

1.3 Tunnuste tüübid

Võimalikke tunnuseid, mille vastu uurijad võivad huvi tunda, on sadu ja tuhandeid. Tunnuse uurimiseks sobivat metoodikat pole tarvis iga tunnuse jaoks uuesti leiutada – on ju võimalik leida keskmist nii hinnetele, saagikusele kui pesakonna suurusele. Kõigi mainitud tunnuste korral sobib keskmise arvutamiseks sama arvutuseeskiri.

Samas pole võimalik leida mullatüüpide keskmist – sellisel näitajal lihtsalt puuduks tähendus.

Kas oleks võimalik jagada tunnuseid selliselt, et ühte gruppi sattunud tunnused (sama tüüpi tunnused) on analüüsitavad kasutades sarnaseid statistikameetodeid? Selgub, et tunnuste taoline jagamine on täiesti võimalik.

Definitsioon 1.3 ***Pideva tunnuse*** *võimalike väärtuste arv on lõpmatu ja iga kahe võimaliku pideva tunnuse väärtuse vahele mahub alati veel üks võimalik pideva tunnuse väärtus.*

Pidevad tunnused on näiteks taime pikkus, looma kaal, temperatuur, fosfaatide kontsentratsioon vees, saagikus, ... Oleks soovitav, et kõik pideva tunnuse väärtused oleksid mõõdetud sama täpsusega (kõik pikkused mõõdetud millimeetri täpsuseni, kõik kaalumistulemused kirja pandud kg täpsusega jne). Igal juhul tuleb aga jälgida, et sama tunnuse kõik väärtused oleksid kirja pandud samades ühikutes (näiteks kilogrammides). Kui ühe elevandi kaaluna läheb kirja number 5300 (kg) ja teise elevandi kaaluks tuleb 4,9 (tonni), siis vaadeldud elevantide keskmiseks kaaluks annaks arvuti 2602,45, millisel numbril muidugi puudub igasugune sisu.

Definitsioon 1.4 ***Diskreetse tunnuse*** *väärtused saavad olla vaid täisarvulised. Peaaegu alati on diskreetse tunnuse väärtused tekkinud millegi loendamisel.*

Diskreetsed tunnused on näiteks pesakonna suurus, terade arv viljapeas, liikide arv ruutmeetril, looma poolt elu jooksul sünnitatud laste arv, ...

Definitsioon 1.5 ***Järjestustunnus*** *on tunnus, mille kõik võimalikud väärtused on järjestatavad.*

Järjestustunnused on näiteks eksperdi hinnang mullaniiskusele (väga kuiv - kuiv - paras - niiske - liigniiske); eksperdi hinnang looduskooslusele (riikliku kaitse alla võtta/ kohaliku kaitse alla võtta/ jätta juhuse hooleks/ buldooseritega hävitada); jälgija hinnang looma agressiivsusele (õel/ kurjavõitu/ normaalne/ rahumeelne/ tuim); aga samuti näiteks haridus, mõõdetuna skaalal algharidus - keskharidus - kõrgharidus - doktorikraad; jne.

Järjestustunnused tekivad sageli subjektiivsete hinnangute andmisel. Hinnangu andmise kriteeriumid võivad hindajati tugevalt erineda ning see võib paratamatult raskendada ka tulemuste hilisemat interpreteerimist. Seetõttu oleks tungivalt soovitav, et kõik hindajad ja ka analüüsi tulemuste hilisemad kasutajad mõistaksid võimalikult ühtemoodi seda, millal mulda on näiteks peetud "väga kuivaks" või kuna peeti kutsut õelaks.

Definitsioon 1.6 ***Nominaalne tunnus*** *on tunnus, mille väärtused pole sisuliselt järjestatavad.*

Nominaalsed tunnused on näiteks sugu, kasvukoht, liik, karvavärvus, lemmikroog, ...

Juhul, kui (nominaalsel) tunnusel on vaid kaks võimalikku väärtust, näiteks nagu tunnusel *SUGU*, siis kutsutakse vastavat tunnust ka **binaarseks** või **dihhotoomseks** tunnuseks.

Pidevaid ja diskreetseid tunnuseid kutsutakse vahel ka arvulisteks (kvantitatiivseteks) tunnusteks ja järjestus- ning nominaalseid tunnuseid kutsutakse mittearvulisteks ehk kvalitatiivseteks tunnusteks.

Näide 1.2 *Igal uuringusse kaasatud liblikal paluti ühe päeva jooksul muneda nii palju mune kui ta jaksab. Liblikat ja tema poolt munetud munasid uuriti põhjalikult. Kogutud andmed on esitatud tabelis 1.3. Märkus: toodud andmed on illustratiivsed ja ei baseeru tegelikel mõõtmistulemustel.*

Tunnused A ja C on pidevad, B on diskreetne, D on järjestustunnus (kasutatud kodeering: 1-väga räbaldunud; 2-kulunud; 3-peaaegu uus; 4-veatu), E on nominaalne (kasutatud kodeering: 1-rohetäpik; 2-suur-pärlmuttertäpik; 3-väike-pärlmuttertäpik).

Tabel 1.3: Liblikad

A liblika suurus	B munade arv	C munetud munade keskmine kaal	D liblika ilu	E liik
11	20	1,3	2	1
10	34	1,8	3	1
12	67	0,9	4	2
7	10	0,7	2	3
12	0	1,0	1	2

1.4 Tunnuste kodeerimine, puuduvad väärtused

Tunnuse väärtuste sisestamisel arvutisse on sageli mõistlik üks või teine (enamasti pikk) väärtus asendada lühendi ehk koodiga. Näiteks tunnuse *PÜÜGIKOHT* väärtuste sisestamisel võime väärtuse "Elva-Vitipalu maastikukaitseala" asemel sisestada numbri "1" jne. Järjestustunnuse puhul on tunnuse väärtuste kodeerimine numbrite abil enamasti tungivalt soovitav. Sealjuures tuleks jälgida, et koodid säilitaksid väärtuste sisulise järjestuse. Seega ei tohi kasutada kodeeringut:

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

küll aga sobivad kodeeringud

```
1 - hea             1 - paha             1 - hea
2 - ei oska öelda   2 - ei oska öelda    0 - ei oska öelda
3 - paha            3 - hea             -1 - paha
```

Iga veidigi suurema uuringu paratamatuks kaaslaseks on puuduvad andmed. Kas keeldub mõni talunik vastamast mõnele küsimusele, unustas doktorant katselapil ettenähtud ajal mõõtmisi tegemas käia või lasi katses kasutatud valge hiir lihtsalt jalga — tegijal juhtub nii mõndagi. Puuduvad andmed võivad analüüsi käigus palju peavalu põhjustada. Sellegi poolest tasub meeles pidada, et puuduvate andmete lihtsalt "äraunustamine" pole enamasti lahendus ja tekitab tavaliselt rohkem probleeme kui lahendab. Sestap on tungivalt soovitav algandmete sisestamisel sisestada ka need kirjed/objektid, kelle kohta andmed (osaliselt) puuduvad. **Puuduvad väärtused peavad andmestikus olema tähistatud nii, nagu ei tähistata andmestikus midagi muud.** Eriti kergesti võivad puuduva väärtusega segi minna näiteks tegelikud mõõdetud väärtused "0" või "vaadeldud omadust ei esinenud".

Näiteks parasiit A olemasolu mõõtev tunnus võiks olla kodeeritud järgmiselt: "1" – parasiit esines; "0" – parasiiti polnud; "." – informatsioon puudub (puuduv väärtus).

1.5 Ülesanded

1. Uuringu käigus koguti andmeid 15 jahimeeste poolt lastud põdra kohta. Kogutud andmed on esitatud tabelis 1.4. Milliseid tunnuseid mõõdeti? Mis tüüpi tunnustega on tegemist? Milline näeks välja objekt-tunnus maatriks antud andmete korral?

2. Loomaembrüodel mõõdeti järgmiste tunnuste väärtused:

   - VANUS1 (vanus päevades)
   - VANUS2 (rakkude pooldumiskordade arv)
   - EMASLOOMA EKSPOSITSIOON ALKOHOLILE (ei/ natuke/ ohtralt)
   - GEENI X MUTATSIOON (esines/ ei esinenud)

Tabel 1.4: Ülesanne 1 - Kolmes metsas lastud põtrade vanused

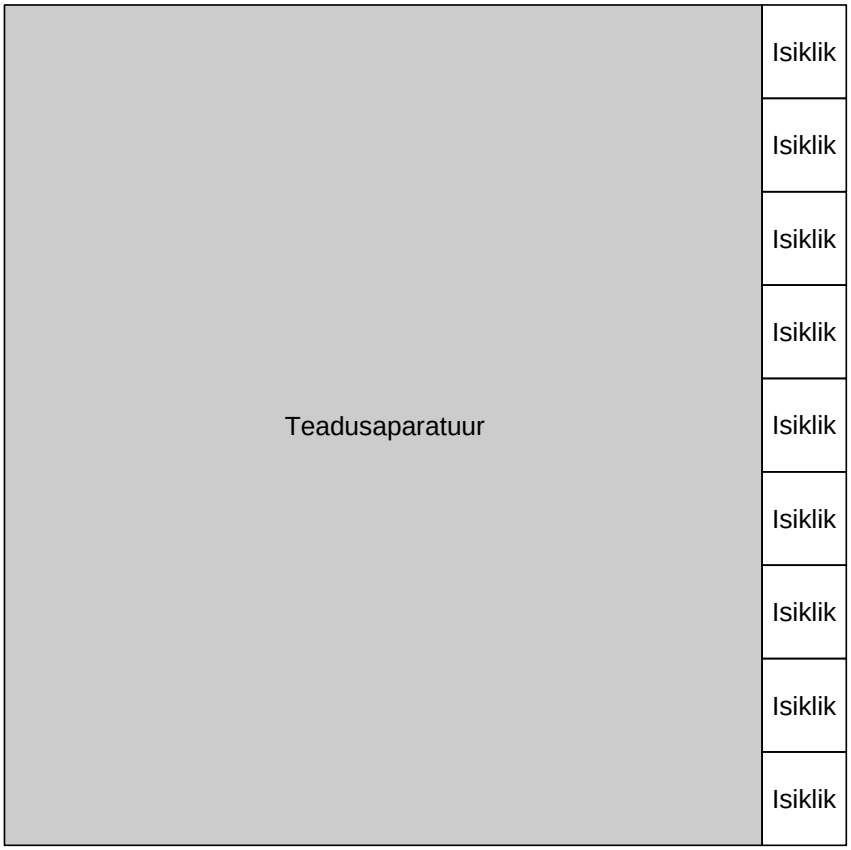
	laskmiskoht		
aasta	Alutaguse	Vändra kant	Kapa-Kohila
2004	10a, 12a	3a	
2005	10a	3a, 5a	7a, 15a, 10a
2006	10a, 11a	4a, 15a	4a, 8a

- KASVUKESKONNA Ph

Mis tüüpi tunnustega on tegemist?

3. Ajakirjanduses ilmusid väited, et Antarktikasse rajatud Eesti uurimisjaama maksumaksja kulul saadetud asjadest on 90% loodusuurijate isiklikud asjad. Loodusuurijad vaidlesid vastu, et isiklikud asjad moodustasid kõnealusest saadetisest vaid 10%. Milles on asi? Kellel on õigus? Vaata ka joonist 1.1!

Joonis 1.1: Saadetis Antarktikasse



Peatükk 2

Kirjeldav statistika

Siin peatükis kuuleme, kuidas saaks lühidalt ja kokkuvõtlikult kirjeldada äraütlemata suurt lasu kokkukogutud andmeid
ehk
ühe tunnuse empiirilise jaotuse kirjeldamine.

2.1 Sagedused ja osakaalud

Kõige lihtsam ja ülevaatlikum viis andmeid kirjeldada, eriti kui tegemist on nominaalse, järjestus- või väheste võimalike väärtustega diskreetse tunnusega on iga võimaliku väärtuse kohta öelda, mitu korda me sellist väärtust oleme näinud (raporteerida iga väärtuse esinemissagedust). Enamasti esitatakse selline kokkuvõtlik informatsioon kas sagedustabelina, tulp- või ringdiagrammina. Vahel on otstarbekam välja tuua erinevate väärtuste osakaalud — näiteks kui suur osa kõigist põldudest asuvad liivastel muldadel, kui suur osa savimuldadel jne. Osakaalusid võib vajaduse korral esitada ka protsentides.

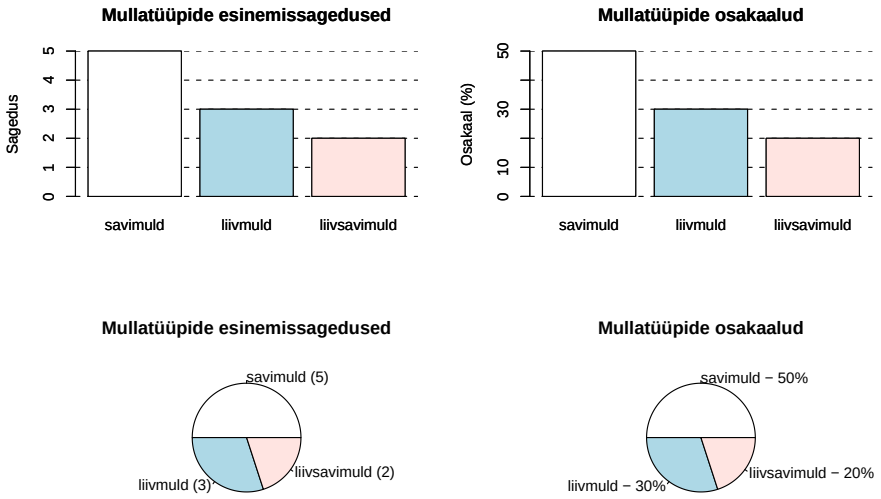
Alljärgnevalt vaatleme näites 1.1 toodud tunnuse *MULLATÜÜP* sagedus- ja jaotustabelit ning nende tabelite põhjal joonistatud tulp- ja ringdiagramme.

Nominaalsete tunnuste väärtuste kirjeldamisel polegi suurt midagi muud võimalik teha, kui esitada tunnuse sagedustabel (või esitada erinevate väärtuste osakaalud). Vahel tuuakse eraldi välja ka tunnuse mood — kõige sagedamini esinenud tunnuse väärtus. Mullatüüpe iseloomustavas näites oleks tunnuse *MULLATÜÜP* moodiks savimuld - savimuldasid esines vaadeldud põldude seas kõige rohkem.

Pideva tunnuse väärtuste iseloomustamisel pole ülaltoodud viisil koos-

Tabel 2.1: Tunnuse *MULLATÜÜP* võimalike väärtuste sagedused ja osakaalud

Mullatüüp	Sagedus	Osakaal	Osakaal (%)
savimuld	5	0,5	50%
liivmuld	3	0,3	30%
liivsavimuld	2	0,2	20%



tatud sagedustabelist aga eriti abi — tabel tuleks liiga pikk ning poleks märkimisväärselt targem kui lihtsalt kõigi tunnuse vaadeldud väärtuste esitamine. Koostamaks mõistlikku sagedustabelit pideva tunnuse tarvis, jagatakse pideva tunnuse väärtused eelnevalt samapikkadeks vahemikeks (Näiteks [0..10),[10..20), jne). Seejärel vaadatakse, kui sageli pideva tunnuse väärtus sattub ühte või teise vaatlusalusesse vahemikku. Mitut vahemikku kasutada? Kindlat reeglit siin pole. Üks soovitus võiks olla järgmine: leia ruutjuur vaatluste arvust. Vali vahemike arvuks mõni täisarv, mis oleks ligikaudu samasuur kui leitud ruutjuur. Näiteks, kui tehtud on 10 vaatlust, siis võiks pideva tunnuse väärtused jagada 3 või 4 vahemikku. Antud reegel on vaid soovitusliku väärtusega, vajadusel võib kasutada ka rohkemaid või vähemaid

vahemikke.

Pideva tunnuse sagedustabeli illustreerimisel kasutatavat joonist kutsutakse histogrammiks. Kui tulpdiagramm oli antud andmete (vaatlustulemuste) korral üheselt määratud, siis samade andmete põhjal võime saada üsnagi erineva kujuga histogramme. Muutes sagedustabeli koostamisel kasutatud vahemikke muutub enamasti ka sagedustabel ja tema põhjal joonistatud histogrammi kuju.

Näide 2.1 *Näites 1.1 on antud suvenisu viljakused erinevatel põldudel. Suvenisu viljakus on pidev tunnus, tema väärtuste jaoks sagedustabeli koostamisel võiksime jagada viljakusandmed näiteks nelja vahemikku. Saadud sagedustabel on antud tabelis 2.2. Antud andmete illustreeriva histogrammi allosas on ära märgitud väikeste joonekestega ka tegelikud vaatlusandmed. Kasutatud vahemike korrektsel kirjeldamisel võib kasutada ka nurk- ja ümarsulge – piiripeale jääv vaatlus pannakse siis kirja sinna vahemikku, kus vastav väärtus piirneb nurksuluga. Seega mõõtmistulemus* 4500 *läheb kirja vahemikku* $[4500 \ldots 5000)$ *ja mitte vahemikku* $[4000 \ldots 4500)$.

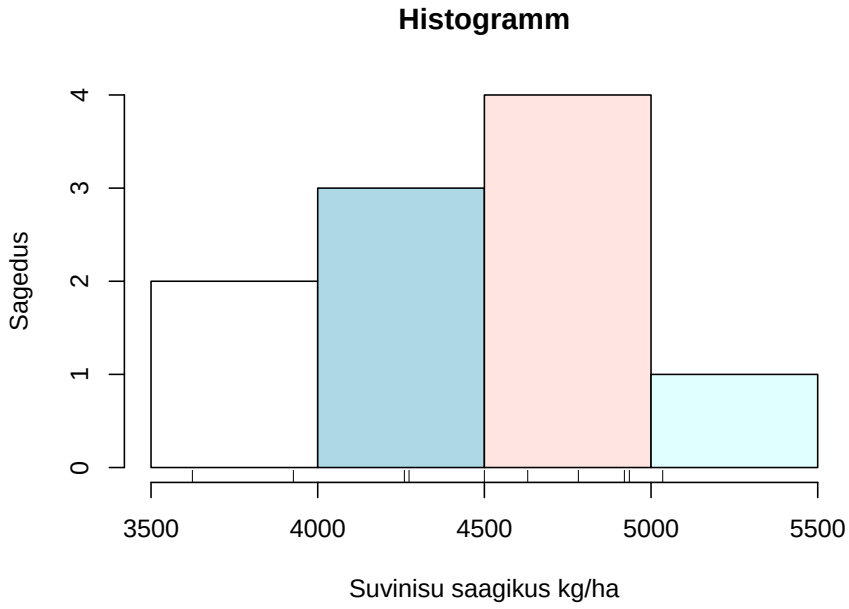
Tabel 2.2: Tunnuse *VILJAKUS* sagedustabel ja osakaalud

Viljakus	Sagedus	Osakaal	Osakaal (%)
$[3500 \ldots 4000)$	2	0,2	20%
$[4000 \ldots 4500)$	3	0,3	30%
$[4500 \ldots 5000)$	4	0,4	40%
$[5000 \ldots 5500)$	1	0,1	10%

NB! On tungivalt soovitav, et kõik kasutatud vahemikud oleksid võrdse pikkusega! Sestap tuleks võimaluse korral vältida ka "avatud" vahemikke, nagu näiteks "suurem kui 50 ha". Kasutades muutuva pikkusega vahemikke võib statistika tarbijas tekitada just sellise pettekujutelma nagu keegi parasjagu soovib.

Näide 2.2 *Kasutades muutuva pikkusega vahemikke sagedustabeli koostamisel võib hooletut või kehva ettevalmistusega statistika tarbijat petta andmeid võltsimatta just nii, nagu parasjagu tarvis. Graafikul 2.2 on samad pideva tunnuse väärtused esitatud kahel erineval moel - kasutades võrdse pikkusega vahemikke sagedustabeli koostamisel (soovitatav tegutsemisviis) ja kasutades muutuva pikkusega vahemikke (petturlus pole haritud inimesele sobiv tegutsemisviis).*

Joonis 2.1: Näites 2.1 toodud sagedustabeli põhjal joonistatud histogramm

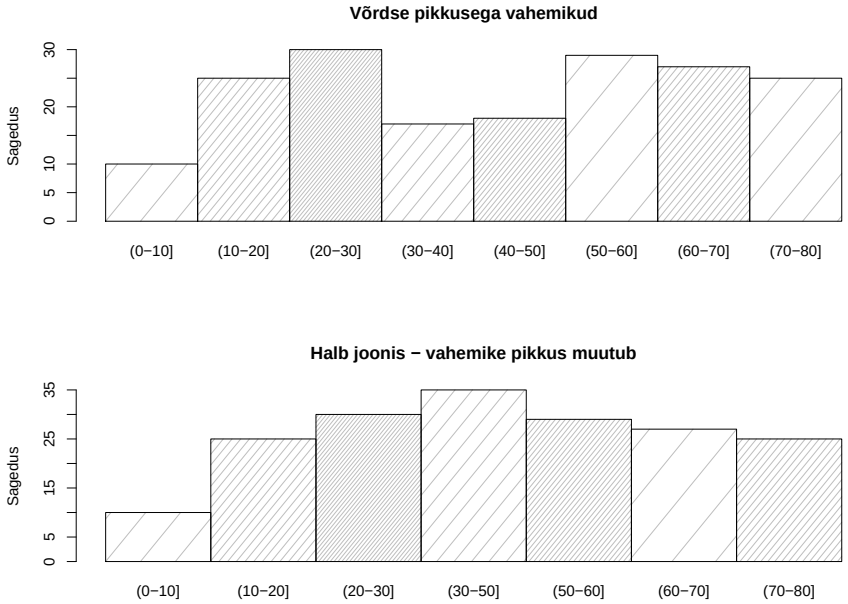


NB! Joonisele tuleb kanda ka vahemikud, kuhu ühtki objekti ei sattunud (kui mingisse vahemikku ei sattunud ühtegi objekti, jätavad paljud programmid sagedustabeli koostamisel vastava rea tabelist välja ja eksitus graafiku joonistamisel on siis juba kerge tulema)!

2.2 Statistikud

Sagedustabel on kasulik viis uuritud tunnuse väärtuste iseloomustamiseks, aga vahel soovime tähelepanu juhtida mõnele meid kõige enam huvitavale andmetega seotud küljele või rõhutatult välja tuua meie andmete omapära. Sealjuures on meile tihti abiks statistikud. *Statistik* on andmete põhjal üheselt arvutatav (enamasti arvuline) näitaja. Arvatavasti tuntuim statistik on keskmine (kõik me oleme muretsenud oma keskmise hinde pärast või lugenud ajalehtedest keskmise palga muutumisest).

Joonis 2.2: Samad andmed - võrdse pikkusega vahemikud ja muutuva pikkusega vahemikud



2.3 Vaatlustulemuste suurust iseloomustavad statistikud

2.3.1 Keskmine

Inglise keeles *mean* või *average*, $\bar{x}$ - uuritava tunnuse väärtuste aritmeetiline keskmine:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right).$$

Näide 2.3 *Eesrindlik talunik Sauna Mats hakkas oma tiigis krokodille kasvatama. Kasvatas neid veidi ja mõõtis siis kõigi oma kuue kasvandiku pikku-*

sed ära: 2,3m 1,7m 0,6m 0,8m 1,4m 2,2m. Nende keskmine pikkus on seega

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}\left(x_1 + x_2 + \ldots + x_n\right) \\ &= \frac{1}{6}\left(2,3 + 1,7 + +0,6 + 0,8 + 1,4 + 2,2\right) = \frac{1}{6}9 = 1,5\,(m).\end{aligned}$$

Keskmise omadusi:

1. $\overline{cx} = c\bar{x}$, kus c on konstant.
   Üks järeldus sellest omadusest: Kui oleksime näiteks samade objektide pikkuseid mõõtnud meetrites (x) ja sentimeetrites ($100x$), siis sentimeetrites tehtud mõõtmiste keskmine ($\overline{100x}$) tuleks sama kui meetrites tehtud mõõtmiste keskmine ($\bar{x}$) korda 100.

2. $\overline{x + c} = \bar{x} + c$, kus c on konstant.
   Kui näiteks peale püütud loomade kaalumist selgus, et kaal polnud täpselt tasakaalus - igal kaalumisel näitas kaal c kilogrammi rohkem, siis valede kaalumiste keskmist $\overline{x + c}$ teades saame arvutada korrektse kaaluga tehtud kaalumiste keskmise: $\bar{x} = \overline{x + c} - c$ ning seega pole keskmise arvutamiseks tarvis mõõtmiseid uuesti teha.

3. $\overline{x + y} = \bar{x} + \bar{y}$
   Üks tudeng arvutas, kui palju kasvavad keskmiselt taimekesed hommiku ja lõuna jooksul ($\bar{x}$). Teine tudeng leidis õhtu ja öö jooksul toimunud juurdekasvude keskmise ($\bar{y}$). Professor leidis aga taimede keskmise ööpäevase juurdekasvu ($\overline{x + y}$) liites oma kahe tudengi tulemused ning kirjutas selle kohta artikli ise ühtegi mõõtmist tegemata.

4. $\sum_{i=1}^{n} x_i = n\bar{x}$
   Triviaalne, kuid igal juhul meelespidamist vääriv. Näiteks karja summaarne väljalüps on leitav korrutades keskmise väljalüpsi karja suurusega.

Dihhotoomse (kahe võimaliku väärtusega) tunnuse keskmine väärib eraldi ära märkimist. Kui meil on näiteks tegemist tunnusega võimalike väärtustega 0 (parasiite pole) ja 1 (parasiite leidub), siis taolise tunnuse keskmiseks tuleb 1-tede osakaal (või, korrutatult sajaga, protsent). Seega on taolisel viisil kodeeritud andmete korral keskmise leidmine kerge viis parasiitidega nakatunud isendite protsendi arvutamiseks.

Keskmine pole alati parim näitaja iseloomustamaks uuritava tunnuse väärtuste suurust. Nimelt on aritmeetiline keskmine tundlik üksikute suurte

väärtuste suhtes - piisab ühestainsast teistest märgatavalt erinevast vaatlusest, et keskmist tugevalt muuta. Seda illustreerib ka järgmine näide.

Näide 2.4 *Uurides maduusside kaalu, saadi vaatlustulemusteks 2,2kg 2,6kg 2,8kg 2,4kg 10,0kg. Viimane madu on sedavõrd kaalukas, kuna on vahetult enne ülekaalumist nahka pistnud saaklooma.*

$$\bar{x} = \frac{1}{5}(2,2 + 2,6 + 2,8 + 2,4 + 10,0) = 20/5 = 4\,(kg).$$

Selgub, et keskmine kaal tuleb suurem kui enamike usside kaal ja peegeldab tugevalt ebatüüpilise, äsjatoitunud looma kaalu. Kuna keskmine võib olla suurem (või väiksem) kui enamik vaatlustulemusi, tekib lahknevus aritmeetilise keskmise kui näitaja ja keskmise intuitiivse tähenduse vahel (intuitiivselt on ju selge, et "keskmine" madu kaalub vähem kui 4 kg!!). Pakkumaks välja teist matemaatilist näitajat, mis iseloomustab andmete "keskmist" suurust sageli intuitiivselt täpsemalt, on kasutusele võetud mediaan.

2.3.2 Mediaan

inglise keeles *median* (või lühendina *med*) - vaatlustulemus, millest suuremaid ja väiksemaid väärtuseid on samapalju. Mediaani leidmiseks järjestatakse kõik vaatlustulemused, saades nn. variatsioonrea: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, kus $x_{(1)}$ on kõige väiksem vaatlustulemus, $x_{(2)}$ on suurem kui $x_{(1)}$ kuid väiksem kõigist ülejäänutest jne kuni vaatlustulemuseni $x_{(n)}$, mis on suurem kõigist teistest. Selle järjestatud vaatlustulemustest moodustatud rea keskel asuv vaatlustulemus ongi mediaan. Kui keskmist vaatlustulemust ei saa leida (kui vaatlustulemusi on paarisarv tükki) siis sobib mediaaniks mistahes arv kahe variatsioonrea keskmise elemendi vahel. Kokkuleppeliselt loetakse sellisel juhul mediaaniks kahe variatsioonrea keskel asuva vaatlustulemuse aritmeetilist keskmist. Matemaatiliselt korrektselt kirjapandult: Kui vaatlusi on tehtud paaritu arv kordi, $n = 2k+1$, siis on vaatlustulemuste mediaaniks variatsioonrea $(k+1)$. element $x_{(k+1)}$. Kui vaatlustulemusi oli aga paarisarv, $n = 2k$, siis loetakse vaatluste mediaaniks variatsioonrea k. ja $(k+1)$. elemendi aritmeetilist keskmist: $med(X) = (x_{(k)} + x_{(k+1)})/2$. Variatsioonrea j-ndat elementi, $x_{(j)}$, nimetatakse j.-ndaks järkstatistikuks. Vaatlustulemuse järjekorranumbrit variatsioonreas nimetatakse astakuks.

Näide 2.5 *Leiame eelmises näites toodud maduusside kaalu mediaani. Esmalt järjestame vaatlustulemused leidmaks variatsioonrida ja saame: 2,2 2,4 2,6 2,8 10,0. Kuna vaatlusi on paaritu arv, $n = 5 = 2\cdot 2+1, k = 2$, siis saame*

mediaaniks variatsioonrea 3. elemendi $med(X) = x_{(2+1)} = x_{(3)} = 2,6$. *Tulemus iseloomustab maduusside harilikku kaalu paremini kui keskmine kaal, sest on vähem mõjutatav eriliste üksikisendite (üksikvaatluste) poolt.*

Mediaani omadustest:

erinevatel põhjustel vaatlusandmeid vahel teisendatakse, näiteks logaritmitakse. Juhul, kui kasutatav teisendus ei muuda vaatluste järjekorda (suurim jääb suurimaks jne — või täpsemalt öeldes: kui kasutatav teisendus on monotoonne), siis võime teisendatud andmete mediaani leida tehes esialgsete andmete põhjal leitud mediaaniga sama teisenduse (näiteks logaritmime). Matemaatilisemas keeles öeldult: teostades mistahes andmete monotoonse teisenduse $f(x)$, st asendades vaatlusandmed $x_1, x_2, \ldots, x_n$ teisendatud vaatlusandmetega $x_1^{uus} = f(x_1), \ldots, x_n^{uus} = f(x_n)$ võime teisendatud andmete mediaani arvutada esialgsete andmete mediaani kasutades:

$$med(x^{uus}) = f(med(x)).$$

Samuti tuleks meeles pidada, et mediaani ja vaatluste arvu teades ei saa välja rehkendada vaatluste summat — mis on tuntav puudus. Riigi mediaanpalka ja töötegijate arvu teades pole riigiametnikul või ärimehel võimalik leida summaarset palkadena ringlevat rahasummat; päevaste läbimüükide mediaani teades ei saa poodnik leida kuu summaarset läbimüüki jne.

Teatavatel juhtudel võib mediaan osutuda liiga "tuimaks" statistikuks: kuigi andmed muutuvad küllaltki palju, mediaan ei muutu.

Näide 2.6 *Soovime võrrelda linnamüras ja saastas elava linnupaari kurna suurust metsarahus pesitseva linnupaari omaga. Kogutud andmed on järgmised:*

Linnas pesitsevad linnud: 1, 1, 1, 2, 2, 2, 2

Metsas pesitsevad linnud: 2, 2, 2, 2, 3, 5, 7

Mõlemal juhul tuleb pesakonna suuruse mediaaniks 2 linnupoega. Seega neid kahte gruppi mediaani abil võrreldes me ei märkakski pesitsedukuse erinevust.

Kuna viimases näites ülestõstetud "liigse-tuimuse-probleem" esineb eelkõige järjestus- või diskreetsete tunnuste korral, tasub mediaani kasutada nimetatud tüüpi tunnuste korral üsna ettevaatlikult.

2.3.3 Mood

Inglise keeles *mode*, lühendina ka *mod* — vaatlustulemus, mida esineb kõige rohkem — väärtus, mis on parajasti moes.

Näide 2.7 *Uuriti metsatukas kasvavate seente liigilist koostist. Saadi tulemuseks: puravik, sitaseen, kukeseen, puravik, kukesen, kukeseen, sitaseen, kukeseen.*

mod(X)=kukeseen.

Arvuliste väärtustega tunnuste jaoks moodi leides võime sattuda olukorda, kus iga või enamik vaatlustest teineteisest erinevad (kui mõõta piisavalt täpselt, selgub, et iga maduussi kaal on teistest erinev). Enamasti kasutatakse siis sagedustabeli abi (vaata sagedustabeli koostamist pidevale tunnusele) ja vahemikku, mis osutus kõige populaarsemaks, loetakse moodiks. Sõltuvalt histogrammi kujust räägitakse vahel ka kahemodaalsest jaotusest (histogrammil on kaks eraldipaiknevat tippu), või multimodaalsest jaotusest (rohkem kui kaks eraldipaiknevat tippu).

Küsimus:
Eksperimendi keskmine kestvus on 4 päeva. Kas reserveerides endale aega eksperimendi läbiviimiseks 5 päeva, võime olla kindlad, et jõuame selle ajaga tulemusteni? Mida oleks vaja teada lisaks antud eksperimendi kestvuse kohta, et suudaksime sellele küsimusele vastata?

2.4 Vaatluste hajuvus

Mõnikord on kõik vaatlused igavalt üheülbalised, teinekord on aga iga uus mõõtmistulemus teistest sedavõrd erinev, nagu polekski mõõdetud ühe ja sama tunnuse väärtust. Kui erinevad teineteistest vaatlustulemused antud tunnuse korral võivad olla?

2.4.1 Miinimum ja maksimum

inglise keeles *maximum*, *minimum*, lühendina kui min, max — sageli kasutatavad ja intuitiivselt hästi mõistetavad tunnuse hajuvust (võimalikku varieeruvust) iseloomustavad statistikud. Teades näiteks eksperimendi minimaalset ja maksimaalset võimalikku kestvust, saaksime kindlalt väita, kas meie poolt eksperimendi jaoks planeeritud 5 päevast piisab.

Tunnuse hajuvuse iseloomustamiseks kasutatakse ka maksimumi ja miinimumi vahet ehk haaret (ka variatsiooniulatus, inglise keeles *range*) — maksimumi ja miinimumi vahe:

$$haare = maksimum - miinimum$$

Ehkki kergesti mõistetavad, esineb miinimumi ja maksimumi kasutamisel ka tõsiseid probleeme. Juhul, kui uuritavaks tunnuseks on jalgade arv küülikul, võime kergesti jõuda tulemuseni: jalgu on küülikul 0 (miinimum) kuni 8 (maksimum). Miks? Sest aeg-ajalt sünnib väärarengutega küülikuid, esineb vigastatud loomi jms. Korraliku uuringu ja ausa uurija puhul kirjeldavad miinimum ja maksimum sageli geneetiliselt muteerunud või muidu väga harukordseid ja erandlikke isendeid või juhtumeid. Tänu omadusele kirjeldada kõige veidramaid juhtumeid, võivad miinimum ja maksimum osutuda praktikas kehvasti kasutatavaks - suur osa vaatlustulemusi on enamasti märksa suuremad kui miinimum ja märksa väiksemad kui maksimum. Praktikas aeg-ajalt esinev lähenemisviis, kus uurija oma meele järgi suvaliselt "ebatüüpiliste" isendite mõõtmistulemused minema viskab enne miinimumi ja maksimumi leidmist, pole teaduskirjanduses lubatav - sest iga uurija jaoks võib "ebatüüpiline" omada erinevat tähendust. See raskendab miinimumi ja maksimumi kasutamist uuritava tunnuse teaduslikul kirjeldamisel, teeb nad aga väärtuslikuks andmetest vigade või veidrike väljaotsimisel.

On ka teine probleem, mis on seotud miinimumi ja maksimumi kasutamisega. Uute mõõtmistulemuste selgumisel saab vaatlustulemuste maksimum ainult kasvada. Näiteks huvitagu meid küülikute kaal. Mõõtnud ära saja või tuhande küüliku kaalud, võib ta ikkagi olla üsna kindel, et kuskil lippab veelgi priskem isend. Sealjuures on üsna võimatu olemasolevate andmete põhjal oletada, kui kaalukas võib olla Eesti Kõige Kaalukam Küülik.

Miinimum ja maksimum on üks võimalus iseloomustada tunnuse hajuvust. Teine võimalus on iseloomustada hajuvust kirjeldades üksikvaatluste kaugust keskmisest. Seda ideed modifitseerides on saadud dispersiooni nime all tuntud statistik.

2.4.2 Dispersioon ja standardhälve

Mõiste dispersiooni vaste inglise keeles on *variance*, tähistus: s^2. Dispersiooni arvutamiseks kasutatakse järgmist valemit:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Üksikvaatluste erinevus keskmisest, $x-\overline{x}$, nimetatakse hälbeks. Dispersiooni saab vaadata kui hälvete ruutude keskmist. Miks dispersiooni arvutamisel keskmise leidmiseks kasutatakse jagajana suurust $(n-1)$ tavalise n-i asemel, sellel peatume lähemalt järgmises loengus.

Kui kõik vaatlused on samasuured (kõigil uuritavatel loomadel on neli jalga), siis on kõik hälbed keskmisest nullid ja uuritava suuruse dispersioon on null (tunnuse "jalgade arv" dispersioon on null). Mida erinevamad keskmisest on vaatlused, seda suurem on ka dispersioon.

valimi standardhälve (s) - inglise keeles *standard deviance*, lühendina ka *sd* või *std*. On samuti tunnuse hajuvust kirjeldav näitaja,

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

Sarnane dispersioonile, kuid on teisendatud viimaks mõõtühikuid võrreldavaks uuritava tunnuse algsete ühikutega. Tunnetuslikult tajutav kui vaatluste teatavat sorti keskmine kaugus keskmisest.

Näide 2.8 *Uuriti kahte gruppi hiiri: metsikuid ja geneetiliselt puhtatõulisi laborihiiri. Mõlemas grupis mõõdeti hiirte reaktsiooni ärritajale. Tulemuseks saadi:*

```
Metsikud hiired: 15, 45, 30, 10, 25
Laborihiired:    20, 25, 30, 25
```

Standardhälbe leidmiseks tuleb esmalt leida mõlema grupi jaoks keskmised:

$$\overline{x_{metsik}} = \frac{1}{5} \times (15 + 45 + 30 + 10 + 25) = 25$$

$$\overline{x_{labor}} = 25$$

Leiame valimi dispersioonid:

$$\begin{aligned} {s_{metsik}}^2 &= \frac{1}{4}((15-25)^2 + (45-25)^2 + (30-25)^2 + (10-25)^2 + (25-25)^2) \\ &= \frac{1}{4}(100 + 400 + 25 + 225 + 0) = 750/4 = 187,5 \\ {s_{labor}}^2 &= \frac{1}{3}((20-25)^2 + (25-25)^2 + (30-25)^2 + (25-25)^2) \\ &= \frac{1}{3}(25 \ + \ 0 \ + \ 25 \ + \ 0) = 50/3 = 16,66... \end{aligned}$$

Kust saame juba valimi standardhälbed mõlema grupi jaoks:

$$\begin{aligned} s_{metsik} &= \sqrt{{s_{metsik}}^2} = \sqrt{187,5} = 13,69\ldots \\ s_{labor} &= \sqrt{{s_{labor}}^2} = \sqrt{16,66\ldots} = 4,08\ldots. \end{aligned}$$

Märkame, et laborihiired reageerivad ärritusele märksa sarnasemalt (neil on väiksem dispersioon ja standardhälve) kui metsikud hiired. Võimalik, et sarnasem reaktsioon on tingitud laborihiirte homogeensemast genofondist.

Standardhälbe ja dispersiooni omadusi: Olgu c konstant ja x uuritav tunnus. Siis

1. $s^2(cx) = c^2 s^2(x)$;
2. $s(cx) = cs(x)$;
3. $s^2(x + c) = s^2(x)$;
4. $s(x + c) = s(x)$;

Teades vaid uuritava tunnuse keskväärtust (populatsiooni keskmist) ja standardhälvet, võime uuritava tunnuse väärtuste kohta öelda järgmist:

- vähemalt $3/4$ uuritava tunnuse väärtustest asuvad keskväärtusele lähemal kui kaks standardhälvet (enamasti asub kahe standardhälbe kaugusel keskväärtusest umbes 95% vaatlustest);
- vähemalt $8/9$ uuritava tunnuse väärtustest asub keskväärtusele lähemal kui kolm standardhälvet (enamasti asub kolme standardhälbe kaugusel keskväärtusest rohkem kui 99% vaatlustest).

2.4.3 Kvantiilid

α-kvantiiliks (α-*quantile*) nimetatakse sellist uuritava tunnuse väärtust, millest väiksemate väärtuste osakaal mõõtmistulemuste seas on α. Näiteks 0,1-kvantiil on selline uuritava tunnuse väärtus, millest väiksemad olid 10% meie mõõtmistulemustest ja 0,5-kvantiil on selline väärtus, millest väiksemaid väärtuseid on 50% (0,5-kvantiil on sama mis mediaan). Lisaks 0,5-kvantiilile kasutatakse sageli ka 0,25-kvantiili ja 0,75-kvantiili, mida kutsutakse ka **alumiseks ja ülemiseks kvartiiliks** (*quartile*).

Olukordades, kus tekib tahtmine kasutada (raporteerida) miinimumi ja maksimumi, soovitatakse kaaluda, kas poleks informatiivsem kasutada mõnda väikest ja suurt kvantiili, näiteks 0,05-kvantiili ja 0,95-kvantiili. Sel viisil

on võimalik vältida mutantide ja andmesisestusvigade eksitavat mõju meid huvitava tunnuse kirjeldamisel ja langeb ära kiusatus andmete paremaks esitamiseks neid võltsida (ebamugavate vaatlus- või katsetulemuste "unustamise" teel).

Kuidas leida kvartiile? Üks traditsiooniline viis on järgmine: ülemise kvartiili hinnangu saame, kui leiame mediaanist suuremate vaatlustulemuste mediaani (variatsioonrea keskelt kuni lõpuni asuvad variatsioonrea elemendid), alumise kvartiiili saamiseks leiame mediaanist väiksemate vaatlustulemuste kvartiili. Kui mediaaniks (mediaani hinnanguks) osutub üks konkreetne variatsioonrea element, siis see vaatlus lisatakse kvartiilide arvutamisel nii mediaanist suuremate kui ka mediaanist väiksemate vaatluste sekka.

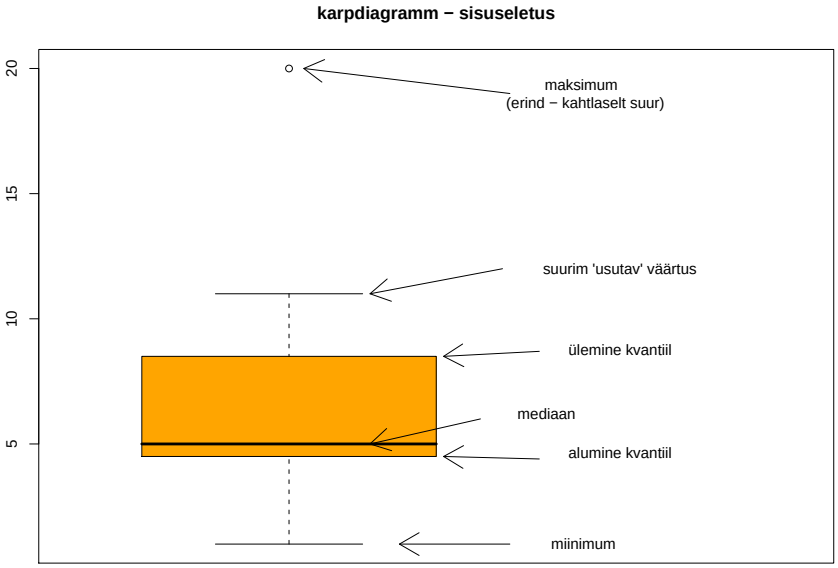
Kuna standardhälve ja dispersioon on (samuti nagu keskmine) tugevalt mõjutatavad üksikute vaatluste poolt, kasutatakse vahel alternatiivina ka kvartiiline vahet iseloomustamaks vaatluste hajuvust.

2.4.4 Karp-vurrud diagramm

Lisaks histogrammile kasutatakse pideva (vahel harva ka diskreetse) tunnuse jaotuse iseloomustamiseks ka **karp-vurrud** diagrammi (inglise k. *boxplot*). Karp moodustub ülemise ja alumise kvartiili vahele, karbi peale märgitakse ka mediaani asukoht(alumine kvartiil oli väärtus, millest väiksemaid väärtuseid oli 25%, ülemine kvartiil oli aga väärtus, millest suuremaid väärtuseid oli 25%). Tekkinud karp "sisaldab" 50% vaatlustulemustest. Lisaks kantakse joonisele suurim ja väikseim vaatlustulemus, valimi miinimum ja maksimum. Miinimumi ja maksimumi ühendamisel karbiga tekivadki nn vurrud. Kui mõni valimis esinevatest uuritava tunnuse väärtustest on väga suur (või väga väike), siis ei tõmmata karpdiagrammi vurre mitte päris selle kahtlaselt suure väärtuseni, vaid mõne veidi pisema (paremini "usutava") väärtuseni. Sellisel juhul kantakse see üks (või enam) "kahtlaselt suurt" väärtust graafikule lihtsalt punktikestena. Kuidas arvuti otsustab, millal on vaatlus kahtlaselt suur (või kahtlaselt väike)? Ühtegi mõistlikku, sisuliselt põhjendatud meetodit sellise otsuse tegemiseks ei kasutata. Võib isegi öelda, et arvuti otsustab lihtsalt oma suva järgi (kuigi mõistagi mingit algoritmi kasutades).

Vaata jooniseid 2.3 ja 2.4.

Joonis 2.3: Karpdiagramm koos selgitustega



Tabel 2.3: Tunnuse tüübile sobivad statistikud

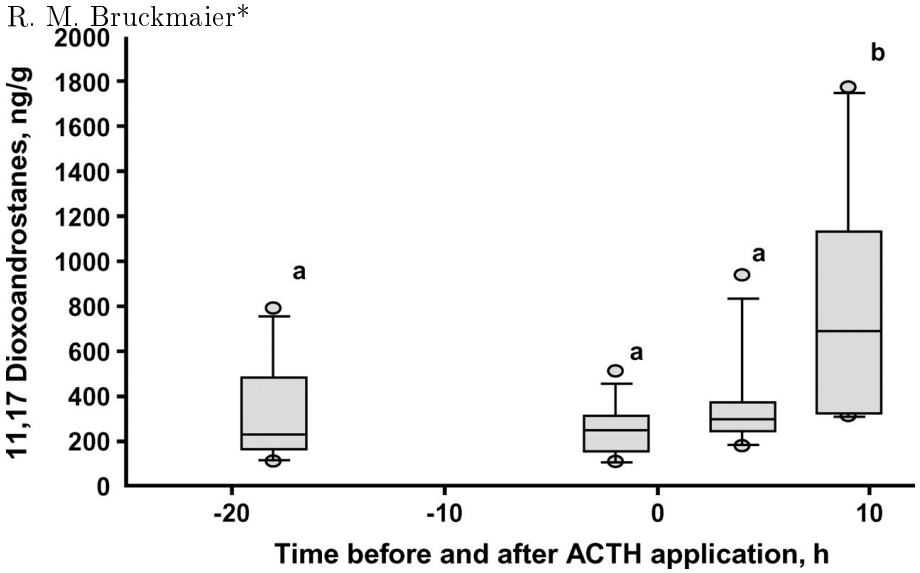
	pidev	diskreetne	järjestustunnus	nominaalne
keskmine	+	+	-	-
mediaan	+	+	+	-
mood	+/-	+	+	+
dispersioon	+	+	-	-
standardhälve	+	+	-	-
kvartiilid	+	+	-	-
histogramm	+	+	-	-
tulpdiagramm	-	+	+	+
karpdiagramm	+	+	-	-

2.5 Teisi statistikuid

Asümmeetriakordaja a:

$$a = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \bar{x})^3,$$

Joonis 2.4: Karpdiagrammi kasutusnäide artiklist J. Anim. Sci. 2004. 82:563-570 Coping capacity of dairy cows during the change from conventional to automatic milking D. Weiss*, S. Helmreich*, E. Möstldagger, A. Dzidic* and R. M. Bruckmaier*



kus s on standardhälve. Sümmeetrilise jaotuse korral on asümmeetriakordaja a väärtus 0: $a = 0$. Kui esineb üksikuid väga suuri väärtuseid (tunnusel on raske saba paremal), siis on asümmeetriakordaja väärtus positiivne. Kui esineb üksikuid väga väikeseid mõõtmistulemusi, siis on asümeetriakordaja väärtus negatiivne. Vaata ka joonist 2.5. Kasutatakse, kui soovitakse rõhutada vaatluste jaotuse sümmeetrilisust/asümmeetrilisust.

Variatsioonikordaja c_v:

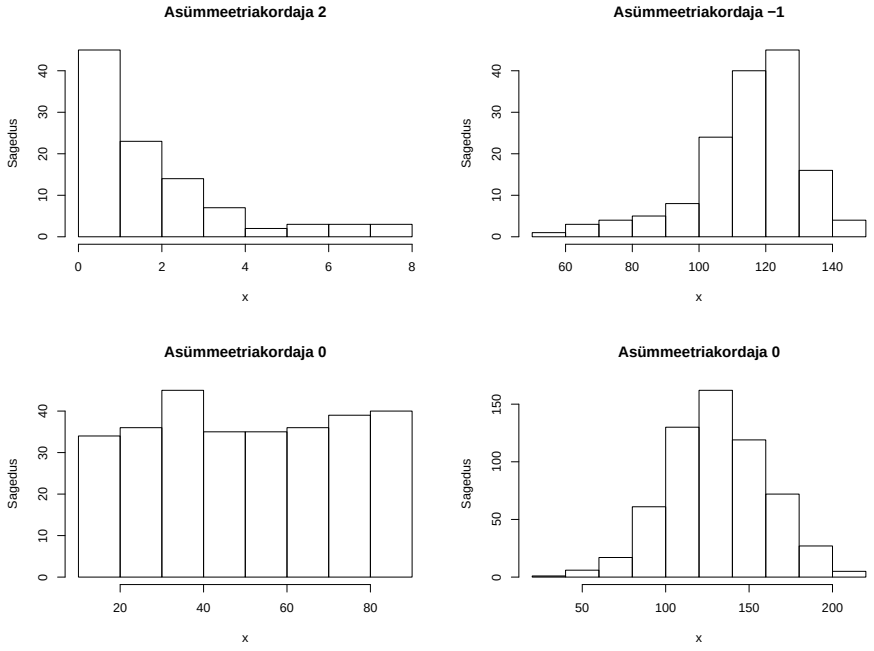
$$c_v = s/\bar{x}.$$

2.6 Ülesanded

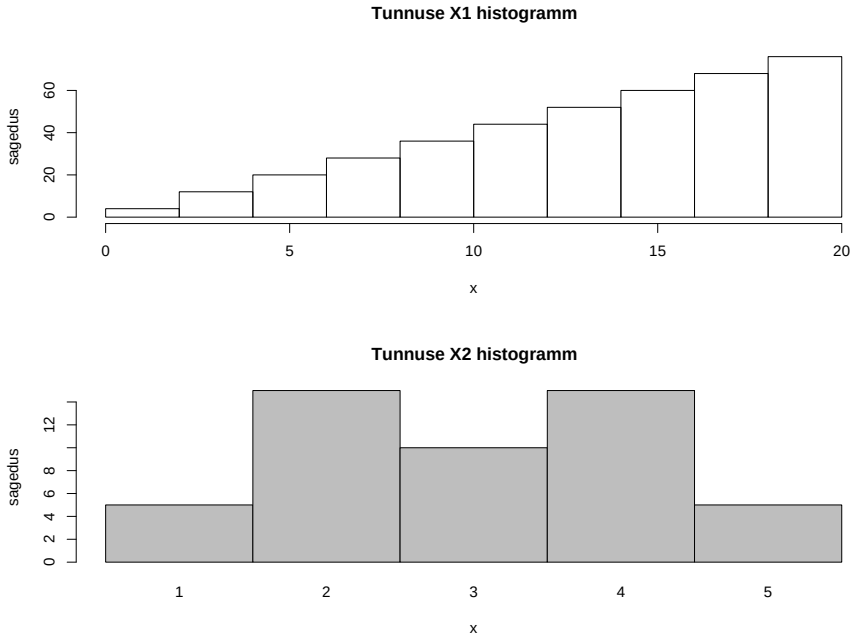
1. Vaata joonisel 1 antud histogramme. Leia nii $X1$ kui $X2$ oletuslik keskmine, dispersioon, mediaan, standardhälve.

2. Joonisel 2 on toodud karpdiagrammid tudengite pikkusele. Kuidas muutub tudengite pikkus sõltuvalt tarbitud õlle kogusest? Millest võiks see olla tingitud?

Joonis 2.5: Asümmeetriakordaja väärtused erinevate jaotuste korral



Joonis 2.6: Ülesanne 1 — milline võiks olla keskmine, standardhälve, mediaan ja dispersioon?



Joonis 2.7: Ülesanne 2. Tartu Ülikooli tudengid ja õlu.