Biomeetria

Statistilise andmeanalüüsi ülesanne. Kirjeldav statistika

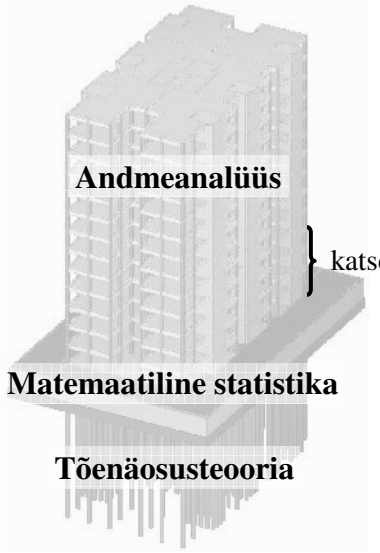
Biomeetria

Biomeetriaks nimetatakse
matemaatika meetodite kasutamist bioloogiliste objektide uurimisel.

(enamasti) ↕

Biostatistikaks nimetatakse
tõenäosusteooria ja matemaatilise statistika meetodite kasutamist
bioloogiliste objektide uurimisel.

Andmeanalüüsi olemus



Andmeanalüüs teeb teaduslikke järeldusi reaalsete (vaatlustest, katsetest, mõõtmistest pärinevate) andmete põhjal, valides rakendatavad statistikameetodid nii, et need võimalikult hästi andmetega sobiksid.

} katseplaan; andmed ja nende esitus; kirjeldav statistika; valim *versus* populatsioon

Matemaatiline statistika tegeleb teoreetiliste andmete $\mathbf{X} = (X_1,\ldots,X_n)$ ja nende funktsioonide $T(\mathbf{X})$ (e statistikute) tõenäosuslike omaduste uurimisega ning statistiliste otsustuste tegemisega.

Statistika ja tõenäosusteooria loomakasvatuses

- populatsioonigeneetika ja evolutsiooniteooria – aitab mõista geneetilise informatsiooni edasikandumist ajas ja ruumis;
- teabe säilitamine, st. arvude kogumine (jõudluskontroll) – kokkuvõtted annavad ettekujutuse loomakasvatuse olukorrast, tasuvusest jne;
- andmetöötlus – jõudlusomaduste, loomakasvatussaaduste kvaliteedi jmt jaotuste ja neid mõjutavate faktorite uurimine, haiguste leviku ja põhjuste uurimine, loomade aretusväärtuste hindamine jne;
- modelleerimiseksperimendid, infotöötlus – uuritakse selektsioonimeetodite efektiivsust, hindamismeetodite täpsust, sööda lõhustuvust, mingi keemilise elemendi segunemist jne;
- teaduslik kirjandus, ajakirjandus – arusaamine esitatud tulemustest (mõisted ja tähistused p, r jne), kriitiline hinnang ajakirjanduses ilmuvate artiklite kohta.

Andmeanalüüsi tüübid

❖ kirjeldav statistika [*descriptive statistics*] – andmete kokkuvõtlik/ülevaatlik esitamine:

- ➢ arvkarakteristikud,
- ➢ sagedustabelid,
- ➢ joonised;

❖ analüüsiv statistika [*inferential statistics*] – andmete põhjal üldiste järelduste ja otsustuste tegemine:

- ➢ parameetrite hindamine,
- ➢ hüpoteeside kontroll,
- ➢ mudelite konstrueerimine.

Objekt

Objekt on uurimisalune ühik, üksikindiviid
(näiteks lehm, talu, põllulapp, firma, inimene, punkt metsas või järvel).

Ka samade andmete puhul võib uurimisobjekti valikuks olla mitu erinevat võimalust.
Näiteks: 2 pesakonda kutsikaid – ühes 2, teises 6 kutsikat

Haiguse levimust uurides võivad uurimispunktideks olla kas lehmad või laudad:

"5% vaadeldud lehmadest põdesid aasta jooksul uuritavat haigust"

vs

"80% vaadeldud lautadest esines uuritavat haigust aasta jooksul".

Objekt – kutsikas		Objekt – pesakond	
Psk nr	Psk suurus	Psk nr	Psk suurus
1	6	1	6
1	6	2	2
1	6		
1	6		
1	6		
1	6		
2	2		
2	2		
Keskmine psk suurus: 5		Keskmine psk suurus: 4	

Lehmade piimajõudlust uurides võib objektideks valida näiteks lüpsikorrad või lüpsipäevad või hoopis laktatsiooni, kusjuures uuritava tunnuse väärtuste stabiilsus võib märgatavalt sõltuda meie valikust (näiteks lüpsikorral lüpstud piimakoguste varieeruvus on ilmselt suurem võrreldes päevalüpside varieeruvusega).

Tunnus

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

Näiteks päevane piimaand, tõug ja vanus lehmi uurides,
talusid uurides talu aastane sissetulek, töötegijate arv, põllumaa pindala ja kaugus lähimast linnast jne.

Et statistika näol on tegu matemaatilise distsipliiniga, ei saa siin kuidagi läbi ilma valemiteta.

Traditsiooniliselt esitatakse tunnuste nimed valemeis suurte tähtedega, näiteks *VANUS*, *TÕUG*, *HAIGUS*. Sageli kasutatakse ka lühendeid – näiteks tunnuse "lehma aastane tingühikutes mõõdetud väljalüps" võime tähistada sümboliga *X*.

Konkreetsete mõõdetud väärtuste tähistamiseks kasutatakse väikeseid tähti ja soovides täpsustada objekti, kellel/millel see väärtus on mõõdetud, esitakse objekti number alaindeksis:
x_3 on tunnuse *X* väärtus 3. objektil (näiteks 3. lehmal).

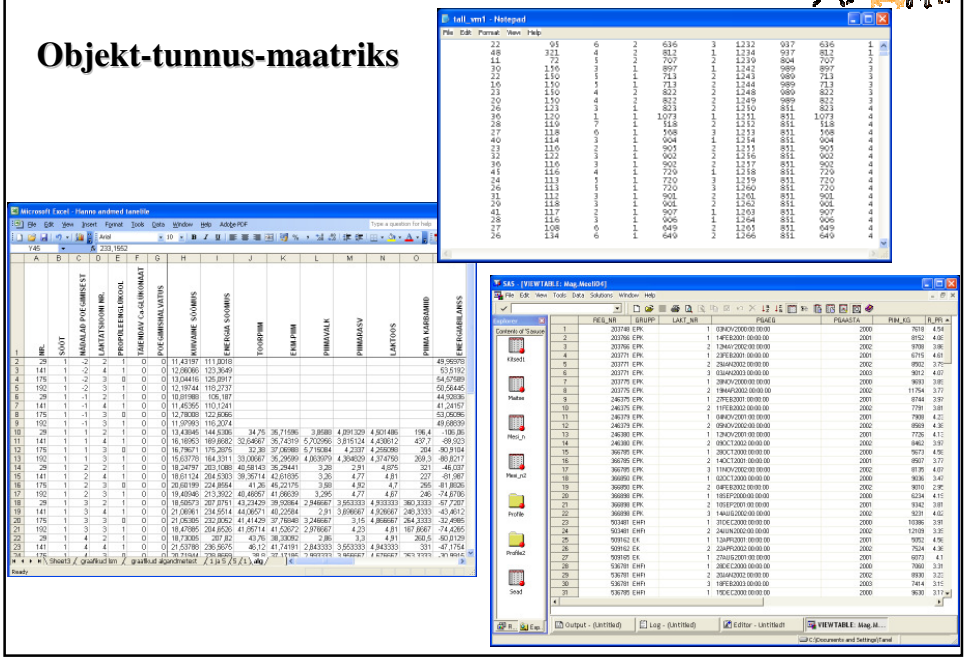
Statistiline andmestik

Objekt-tunnus-maatriks – tabel, kus iga veerg kujutab ühte tunnust ja iga rida ühte objekti.

Näide. Ühest tallist koguti andmed müügis olevate hobuste tõu, värvuse ja hinna kohta. Saadud andmemaatriks (-tabel) on järgmine:

Tõug	Värvus	Hind
tori	raudjas	9000
trakeen	hall	26000
tori	raudjas	16000
eesti hobune	kõrb	9000
eesti hobune	raudjas	12000
trakeen	kõrb	35000
tori	kõrb	17000
eesti hobune	raudjas	14000
tori	must	21000
eesti hobune	kõrb	19000
eesti hobune	must	22500

Objekt-tunnus-maatriks



Tunnuste tüübid

Arvulised e. kvantitatiivsed tunnused [*numerical*]

Diskreetse [*discrete*] **tunnuse** väärtused saavad olla vaid täisarvulised, peaaegu alati on need tekkinud millegi loendamisel.
Näiteks pesakonna suurus, terade arv viljapeas, laktatsiooni number, …

Pideva [*continuous*] **tunnuse** võimalike väärtuste arv lõpmatu ja iga kahe võimaliku väärtuse vahele mahub veel vähemalt üks pideva tunnuse võimalik väärtus; pideva tunnuse väärtused saadakse enamasti millegi otsesel mõõtmisel.
Näiteks piimatoodang, villa pikkus, esmapoegimise iga, saagikus, pH, …

Soovitused:
✓ kõik tunnuse väärtused peaksid olema mõõdetud sama täpsusega,
✓ sama tunnuse väärtuste puhul tuleks kasutada samu ühikuid
(lehma 1 toodang 6300 (kg), lehma 2 toodang 7,9 (tonni) – keskmine toodang 3153,95!?).

Tunnuste tüübid

Mittearvulised e. kvalitatiivsed tunnused [*categorical*]

Järjestustunnuse [*ordinal*] väärtuste vahel on võimalik objektiivne järjestus (hinnangud etteantud skaalal jm).
Näiteks haridus (alg- / kesk- / kõrgharidus / doktorikraad), poegimisraskus, hinnang mulla niiskusele (väga kuiv / kuiv / paras / niiske / liigniiske), hinnang pulli välimusele (niru / normaalne / kaunis), ...

Probleemiks võimalikud subjektiivsed hinnangud (milline pull on kaunis?)!

Nominaalsed tunnused [*nominal*] on mittearvulised tunnused, mille väärtuste vahel ei ole sisulist järjestust.
Näiteks tõug, värvus, farm, kasvukoht, ...

Binaarsed (**dihhotoomsed**) **tunnused** on kahe väärtusega nominaalsed tunnused.
Näiteks sugu.

Tunnuste tüübid

Näide. Uuriti Lõuna-Eestis asuvaid talusid, kogutud andmed on esitatud järgnevas tabelis.

A	*B*	*C*	*D*	*E*
suurus (ha)	müügitulu	peamine tegevusala	põllumaa kvaliteet	talupere suurus
25	340000	1	1	5
15	220000	2	2	4
44	700000	1	3	4
12	500000	3	3	2
20	1200000	2	1	2

Tunnused *A* ja *B* on pidevad, *C* nominaalne (kasutatud kodeering: 1-karjakasvatus; 2-viljakasvatus; 3-turism), *D* on järjestustunnus (kasutatud kodeering: 1-väga hea; 2-keskmine; 3-kehv) ning tunnus *E* on diskreetne tunnus.

Tunnuste kodeerimine

Kodeerimine – sõnaliste vastusevariantide arvudega asendamine.

Näiteks tunnuse "arvamus valitsusest" väärtuste sisestamisel võime vastsevariandi "valitsus on hea" asemel sisestada numbri "1", vastusevariandi "valitsus on kesk-pärane" asemel numbri "2" ja vastusevariandi "valitsus on saast" asemel numbri "3"

❖ Järjestustunnuste kodeerimisel tuleb jälgida, et koodid säilitaksid väärtuste sisulise järjestuse.

Vastuse variandid	Vastuse variantide koodid			
hea	~~1~~	1	3	1
halb	~~2~~	3	1	-1
ei tea	~~3~~	2	2	0

❖ Binaarse tunnuse kodeerimisel on eelistatav lihtsaim võimalus, näiteks 0 ja 1 (või ka 1 ja 2, kui see on sisuliselt mõistetavam).

❖ Nominaaltunnuseid ei ole enamasti vaja arvuliseks kodeerida, ja kui kodeerida, siis koodid sisulist tähendust ei oma (loogiline oleks näiteks järjestada väärtused tähestiku järjekorras).

Statistiline andmestik – märkusi ja soovitusi

Ka hästi planeeritud uurimuse korral võib juhtuda, et kõigi objektide korral ei ole teada kõigi tunnuste väärtusi ja andmestik jääb lünklikuks.

Puuduv väärtus peab olema tähistatud nii, nagu ei tähistata andmestikus midagi muud.

Näiteks parasiit A olemasolu mõõtev tunnus võib olla kodeeritud järgmiselt: "1" – parasiit esines; "0" – parasiiti polnud; "." – informatsioon puudub.

NB! Andmeid *Excelis* analüüsides tuleb puuduvale väärtusele vastav lahter jätta **tühjaks**!

Kirjeldav statistika – sagedused ja osakaalud

Mittearvuliste või diskreetsete tunnuste (erinevate väärtuste arv suhteliselt väike) ülevaatlikuks kirjeldamiseks on lihtne kokku lugeda, mitu korda iga erinevat väärtust esineb, ja esitada saadud arvud tabeli kujul.

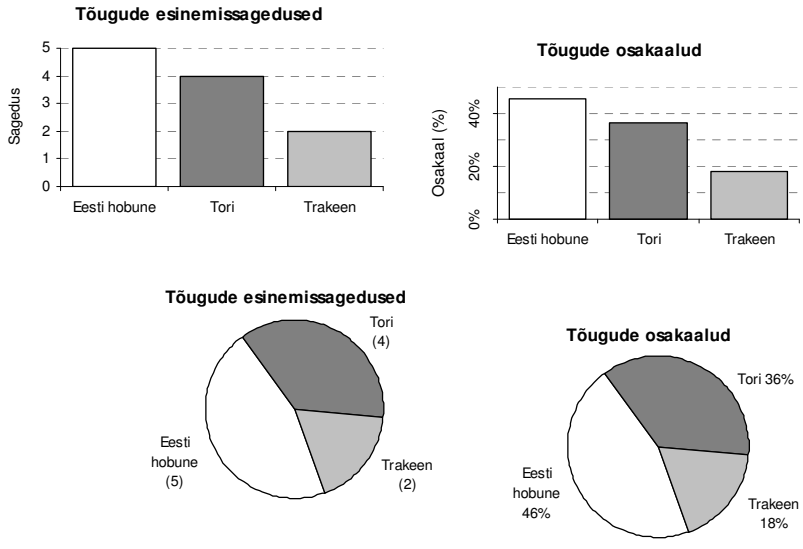
Väärtuse esinemiste arvu nimetatakse tema **sageduseks**.

Tihti leitakse lisaks iga väärtuse (protsentuaalne) **osakaal** valimis, mida nimetatakse ka **suhteliseks sageduseks**.

Tõug	Värvus	Hind
tori	raudjas	9000
trakeen	hall	26000
tori	raudjas	16000
eesti hobune	kõrb	9000
eesti hobune	raudjas	12000
trakeen	kõrb	35000
tori	kõrb	17000
eesti hobune	raudjas	14000
tori	must	21000
eesti hobune	kõrb	19000
eesti hobune	must	22500

Tõug	Sagedus	Osakaal	Osakaal (%)
eesti hobune	5	0,455	45,5%
tori	4	0,364	36,4%
trakeen	2	0,182	18,2%

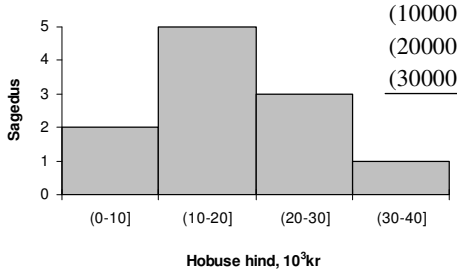
Sagedustabeli asemel võib kokkuvõtliku info väärtuste esinemissagedustest esitada ka kas **tulp-** või **ringdiagrammina** (sektordiagrammina).



Sagedused ja osakaalud

Pidevate tunnuste puhul on tunnuse võimalike väärtuste arv (teoreetiliselt) lõpmatu – seega kui sagedustabelis vastaks igale väärtusele üks rida, siis kaoks praktiliselt erinevus sagedustabeli ja originaalandmete vahel. Seetõttu jagatakse tunnuse võimalikud väärtused intervallidesse ja sagedustabel näitab, mitu väärtust langeb ühte või teise intervalli. Intervallide arv ei tohiks olla liiga suur ja see oleneb valimi suurusest ($\approx\sqrt{n}$).

Hind	Sagedus	Osakaal	Osakaal (%)	Jaotus
(0-10000]	2	0,182	18%	18%
(10000-20000]	5	0,455	45%	64%
(20000-30000]	3	0,273	27%	91%
(30000-40000]	1	0,091	9%	100%

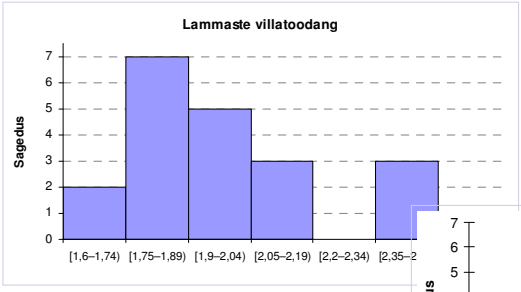


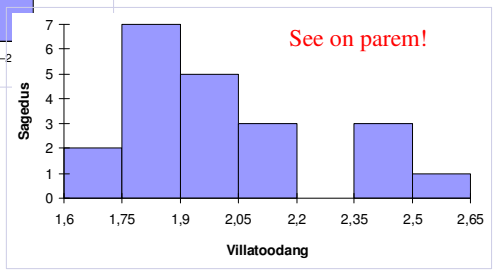
Pideva tunnuse sagedustabeli põhjal saadud tulpdiagrammi nimetatakse **histogrammiks**.

Histogrammid – märkusi ja soovitusi

- Sagedustabelis esitatud intervallide asemel esitatakse histogrammi *x*-telg enamasti pideval skaalal (selle *Excelis* teostamine on üpris keeruline …).

Villatoodang	Sagedus
[1,6–1,74)	2
[1,75–1,89)	7
[1,9–2,04)	5
[2,05–2,19)	3
[2,2–2,34)	0
[2,35–2,49)	3
[2,5–2,64)	1





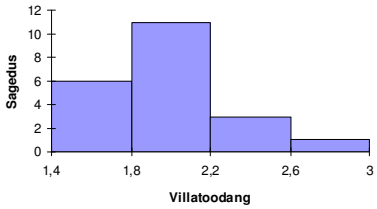
Histogrammid – märkusi ja soovitusi

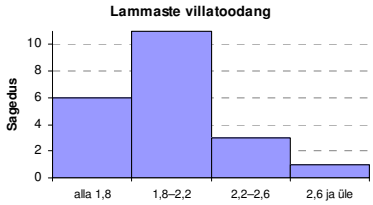
❖ Avatud vahemikke tuleks võimaluse korral vältida.

Villatoodang	Sagedus
alla 1,8	6
1,8–2,2	11
2,2–2,6	3
2,6 ja üle	1

Villatoodang	Sagedus
1,4 – 1,8	6
1,8 – 2,2	11
2,2 – 2,6	3
2,6 – 3,0	1

See on parem!

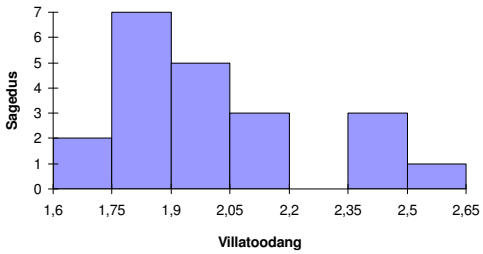


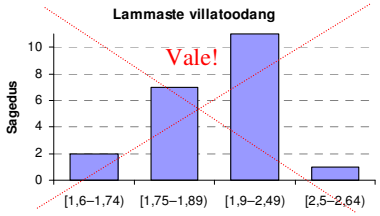


Histogrammid – märkusi ja soovitusi

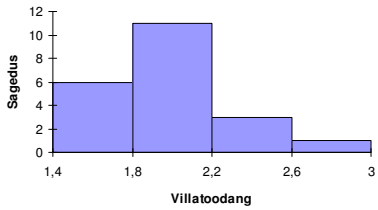
❖ On tungivalt soovitatav, et kõik kasutatud vahemikud oleksid võrdse pikkusega!

❖ Joonisele tuleb kanda ka vahemikud, kuhu ühtki objekti ei sattunud!



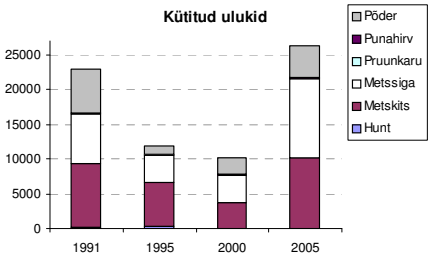


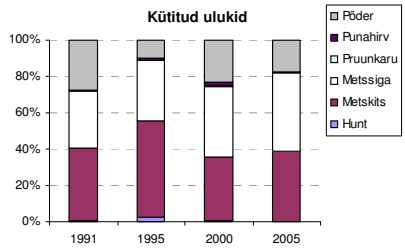
❖ Erinevalt tulpdiagrammist, mis on antud andmete korral üheselt määratud, võime samade andmete põhjal konstrueerida erinevaid histogramme.



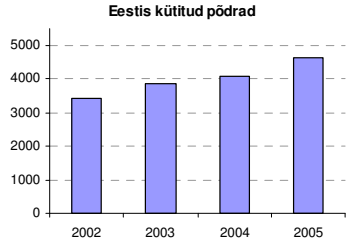
Sagedused ja osakaalud – märkusi ja soovitusi

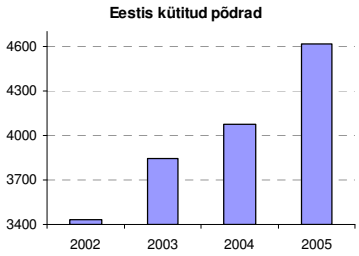
- Sagedused *versus* osakaalud





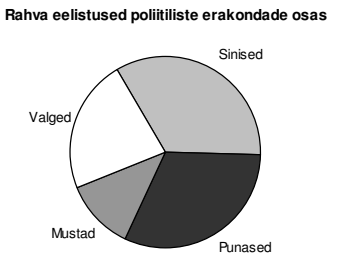
- Skaala!



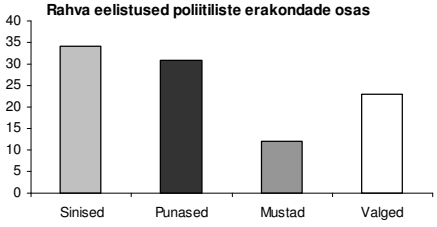


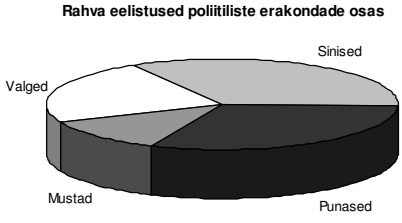
Sagedused ja osakaalud – märkusi ja soovitusi

- Vältida tuleks 3-mõõtmelisi graafikuid, eriti ringdiagramme.



Erakond	Osakaal (%)
Sinised	34
Punased	31
Mustad	12
Valged	23





Kirjeldav statistika – arvkarakteristikud

- andmestiku suurus (**valimi maht**) – n
- (aritmeetiline) **keskmine** [*average*, *mean*] – $\bar{x} = \sum_{i=1}^{n} x_i / n$
- **mediaan** (nn 50%-punkt) [*median*]
- **mood** [*mode*] – enim esinev (suurima sagedusega) väärtus

Näide. Uuringu all olnud 5-l haigestunud loomal määrati haiguse peiteajaks vastavalt 8, 16, 12, 60 ja 14 päeva (üks uuritud loomadest oli ilmselt geneetiliselt erinev või siis sai juba mingit muud, haiguse avaldumist pärssivat ravi). Haiguse keskmine peiteaeg on

$$\bar{x} = \frac{8+16+12+60+14}{5} = \frac{110}{5} = 22 \text{ päeva.}$$

Peiteaeg, millest pooltel loomadel avaldus haigus varem ja pooltel hiljem, on leitav kui kasvavalt järjestatud peiteaegade keskmine väärtus e mediaan:

8, 12, 14, 16, 60
= *med*

Kirjeldav statistika – arvkarakteristikud

- **standardhälve** [*standard deviation*] – $s = \sqrt{1/(n-1) \sum_{i=1}^{n} (x_i - \bar{x})^2}$
- **dispersioon** [(*sample*) *variance*] – s^2
- **standardviga** [*standard error*] – $se = s/\sqrt{n}$
- **variatsioonikordaja** [*coefficient of variation*] – $v = s/\bar{x} \times 100\%$

(omab mõtet üksnes positiivsete väärtustega tunnuste korral!)

Näide. Uuriti 5 metsiku ja 4 puhtatõulise laborihiire reaktsiooni ärritajale. Tulemuseks saadi järgmised väärtused:

metsikud hiired – 15, 45, 30, 10, 25; labori hiired – 20, 25, 30, 25.

Keskmised reaktsioonid kummagi grupi jaoks on

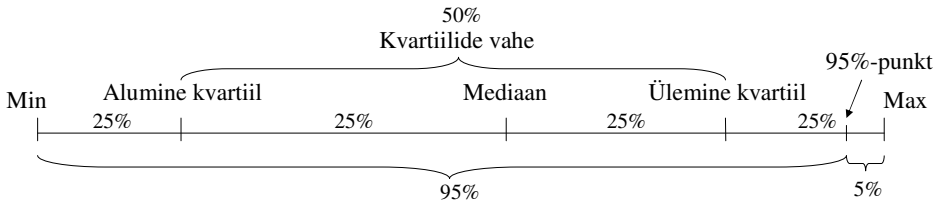
$$\bar{x}_m = \frac{15+45+30+10+25}{5} = \frac{125}{5} = 25, \quad \bar{x}_l = \frac{20+25+30+25}{4} = \frac{100}{4} = 25.$$

$$s_m = \sqrt{\frac{(15-25)^2+(45-25)^2+(30-25)^2+(10-25)^2+(25-25)^2}{5-1}} = \sqrt{\frac{750}{4}} = \sqrt{187{,}5} \approx 13{,}69;$$

$$s_l = \sqrt{\frac{(20-25)^2+(25-25)^2+(30-25)^2+(25-25)^2}{4-1}} = \sqrt{\frac{50}{3}} \approx \sqrt{16{,}67} \approx 4{,}08.$$

Kirjeldav statistika – arvkarakteristikud

- **kvartiilid** – alumine kvartiil e 25%-punkt ja ülemine kvartiil e 75%-punkt
- **kvartiilide vahe** [*lower*, *upper quartile*; *interquartile range*]
- detsiilid, protsentiilid e protsendipunktid**/kvantiilid**
- min, max



Karp-vurrud diagramm