# Statistika ja modelleerimise baaskoolitus õppejõududele ja juhendajatele

**Loeng 1**

Eesti Maaülikool
27.-30. august 2012
Tanel Kaart

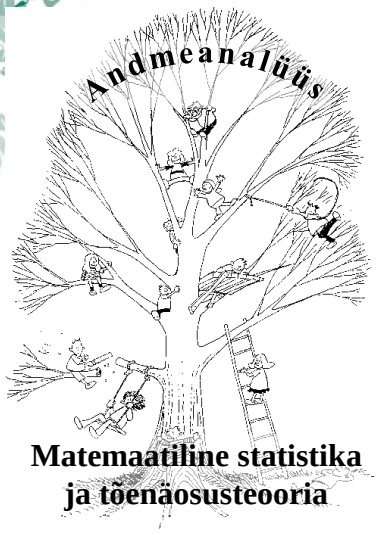
## Koolituse eesmärk

✓ Klassikalise andmeanalüüsi põhimõtete ja -meetodite meenutamine ja tutvustamine.

✓ Tähelepanu on järgmistel küsimustel:

- mis kujul sisestada andmed arvutisse ja millest alustada andmete analüüsil;
- kuidas kirjeldada andmeid ja esitada analüüside tulemusi;
- mis loogikal baseeruvad andmeanalüüsil teostatavad arvutused ja vastuvõetavad otsused;
- kuidas mõõta ja testida tunnuste vahelisi seoseid ning gruppide vahelisi erinevusi
  (korrelatsioon- ja regressioonanalüüs, sagedustabelid, $\chi^2$-test ja Fisheri täpne test, logit- ja probit-regressioon, t-test, mitteparameetrilised testid, dispersioonanalüüs).

## Andmeanalüüsi olemus




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

**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.

## 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.

## Andmeanalüüsi tüübid

✓ Eksperimendipõhine uuring [*experiment-based study*]

- katse-eelselt määratud ja täpselt kontrollitud tingimused,
- väike ja erinevatele (võrreldavatele) tingimustele vastav sama arv vaatlusi/mõõtmisi,
- analüüsimiseks standardsed statistikameetodid.

✓ Mudelipõhine uuring [*model-based study*]

- täpselt ettemääramata tingimustes sooritatud vaatlused/mõõtmised,
- sageli suur, segane ja ebaühtlane andmebaas,
- analüüsimeetodid sõltuvad uurija ettekujutusest uuritavaist suurustest ja neid mõjutavaist tegureist, samuti kogutud andmete hulgast ja struktuurist.

## Andmed

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

## Objekt ja tunnus

**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

Metsa uurides võivad uurimispunktideks olla kas puud või punktid metsas –
"80% vaadeldud puudest olid pajud"
*vs*
"35% vaadeldud metsaaladest olid kaetud pajudega".

| Objekt – kutsikas | |
|---|---|
| Psk nr | Psk suurus |
| 1 | 6 |
| 1 | 6 |
| 1 | 6 |
| 1 | 6 |
| 1 | 6 |
| 1 | 6 |
| 2 | 2 |
| 2 | 2 |
| Keskmine psk suurus: 5 | |

| Objekt – pesakond | |
|---|---|
| Psk nr | Psk suurus |
| 1 | 6 |
| 2 | 2 |
| | |
| | |
| | |
| | |
| | |
| | |
| 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).

## Objekt ja 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,
metsapuid uurides võivad mõõdetavateks tunnusteks olla liik, ümbermõõt, kõrgus, vanus jne.

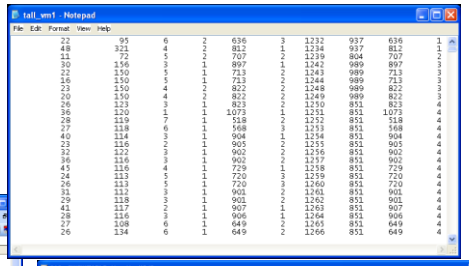
Et statistika näol on tegu matemaatilise distsipliiniga, ei saa siin kuidagi läbi ilma valemiteta.

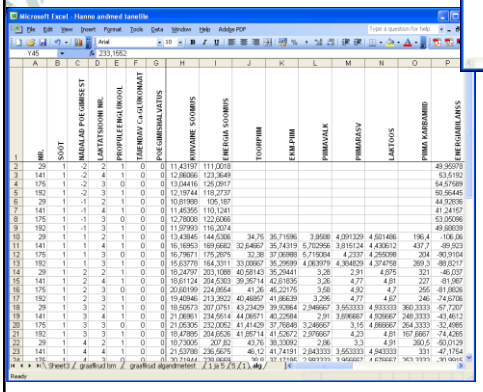
Traditsiooniliselt esitatakse tunnuste nimed valemeis suurte tähtedega, näiteks *VANUS*, *TÕUG*, *SAAGIKUS*. Sageli kasutatakse ka lühendeid – näiteks tunnuse "lehma aastane tingühikutes mõõdetud väljalüps" võime tähistada sümboliga *X*.

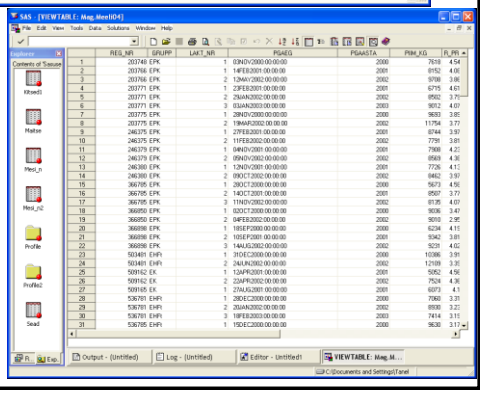
Konkreetsete mõõdetud väärtuste tähistamiseks kasutatakse väikeseid tähti ja soovides täpsustada objekti, kellel/millel see väärtus on mõõdetud, esitakse objekti number alaindeksis:
$x_3$ on tunnuse *X* väärtus 3. objektil (näiteks 3. lehmal).

## Statistiline andmestik

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







## 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
(põllu 1 saagikus 5300 (kg/ha), põllu 2 saagikus 4,9 (tonni/ha) – keskmine saagikus 2602,45).

## 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 kodeerimine

Kodeerimine – sõnaliste vastusevariantide arvudega asendamine.
Näiteks tunnuse "arvamus valitsusest" väärtuste sisestamisel võime

vastusevariandi "valitsus on hea" asemel sisestada numbri "1",
vastusevariandi "valitsus on keskpärane" asemel numbri "2" ja
vastusevariandi "valitsus on saast" asemel numbri "3"

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

✓ 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).

## Puuduvad väärtused

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.

**Plaanides andmeid analüüsida standardse statistikatarkvara (*SAS*, *R*, *Statistica*, …) või mõne tabelarvutussüsteemi (*MS Excel*, *Open Office*, …) abil, on mõistlik jätta puuduvale väärtusele vastav lahter tühjaks.**

## Statistiline andmestik

**Ankeedivastused**

✓ Andmetabeli igasse lahtrisse sisestatakse üks arv või sõna.

| Talu | Tegevusala |
|---|---|
| 1 | karjakasvatus |
| 2 | karjakasvatus, viljakasvatus |
| 3 | viljakasvatus |
| 4 | turism |
| 5 | viljakasvatus, turism |

| Talu | Tegevusala | | |
|---|---|---|---|
| | karjakasvatus | viljakasvatus | turism |
| 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 |
| 5 | 0 | 1 | 1 |

## Statistiline andmestik

**Ajas/ruumis korduvad mõõtmised**

| Nimi | Regnr | Lakt. | 1.seemen-duse aeg | Seemen-duste arv | Aeg1 | Glükoos1 (mg/dl) | Aeg2 | Glükoos2 (mg/dl) | Aeg3 | Glükoos3 (mg/dl) | Aeg4 | Glükoos4 (mg/dl) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALBI | 5030 | 7 | 64 | 6 | -12 | 35,1 | 10 | 28,2 | 37 | 23,7 | 64 | 30,4 |
| SEEVIK | 5383 | 5 | 74 | 5 | -12 | 34 | 10 | 19,7 | 37 | 24,7 | . | . |
| RIIBU | 5537 | 4 | 89 | 1 | -14 | 46,3 | 13 | 23,3 | 26 | 24,9 | 76 | 29,7 |

✓ Kui vähegi võimalik, tuleks mõõtmisi sooritada kõigil objektidel ühesuguste, regulaarsete ajavahemike järel.

(siis pole ka mõõtmise aega näitavaid veerge vaja)

| Nimi | Regnr | Laktat-sioon | Periood | Esimese see-menduse aeg | Seemen-duste arv | Mõõtmis-aeg | Glükoos (mg/dl) |
|---|---|---|---|---|---|---|---|
| ALBI | 5030 | 7 | 1 | 64 | 6 | -12 | 35,1 |
| ALBI | 5030 | 7 | 2 | 64 | 6 | 10 | 28,2 |
| ALBI | 5030 | 7 | 3 | 64 | 6 | 37 | 23,7 |
| ALBI | 5030 | 7 | 4 | 64 | 6 | 64 | 30,4 |
| SEEVIK | 5383 | 5 | 1 | 74 | 5 | -12 | 34 |
| SEEVIK | 5383 | 5 | 2 | 74 | 5 | 10 | 19,7 |
| SEEVIK | 5383 | 5 | 3 | 74 | 5 | 37 | 24,7 |
| RIIBU | 5537 | 4 | 1 | 89 | 1 | -14 | 46,3 |
| RIIBU | 5537 | 4 | 2 | 89 | 1 | 13 | 23,3 |
| RIIBU | 5537 | 4 | 3 | 89 | 1 | 26 | 24,9 |
| RIIBU | 5537 | 4 | 4 | 89 | 1 | 76 | 29,7 |

## Kirjeldav statistika

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

# Sagedused ja osakaalud – diskreetne tunnus

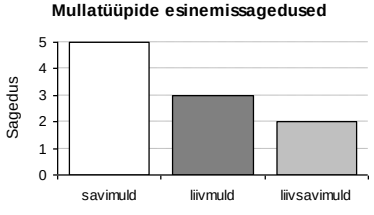
Mittearvuliste või diskreetsete tunnuste (erinevate väärtuste arv suhteliselt väike) ülevaatlikuks kirjeldamiseks on lihtne lugeda kokku, mitu korda iga erinevat väärtust esineb ja kirjutada saadud arvud tabeli kujul. Väärtuse esinemiste arvu nimetatakse tema **sageduseks**.

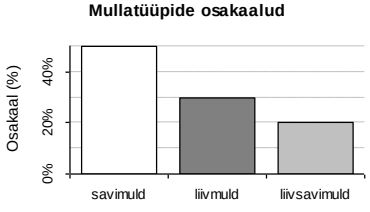
Tihti leitakse lisaks iga väärtuse (protsentuaalne) **osakaal** valimis, mida nimetatakse ka **suhteliseks sageduseks**.

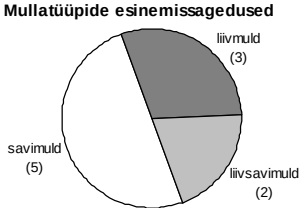
| Mullatüüp | Niiskus | Suvinisu viljakus (kg/ha) |
|---|---|---|
| savimuld | niiske | 3624 |
| liivsavimuld | paras | 4782 |
| savimuld | niiske | 4274 |
| liivmuld | kuiv | 3927 |
| savimuld | paras | 4630 |
| liivmuld | paras | 4920 |
| savimuld | niiske | 4260 |
| savimuld | paras | 4935 |
| liivsavimuld | paras | 5035 |
| liivmuld | kuiv | 4500 |

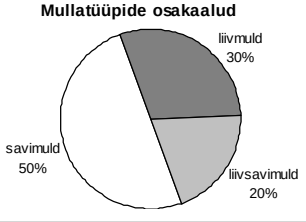
| Mullatüüp | Sagedus | Osakaal | Osakaal (%) |
|---|---|---|---|
| savimuld | 5 | 0,5 | 50% |
| liivmuld | 3 | 0,3 | 30% |
| liivsavimuld | 2 | 0,2 | 20% |

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



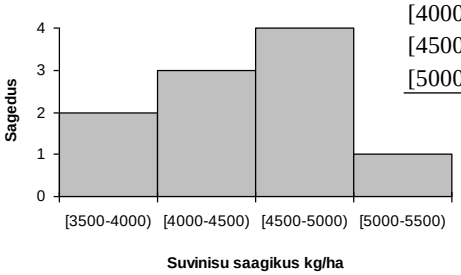



# Sagedused ja osakaalud – pidev tunnus

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 ($\approx\sqrt{n}$).

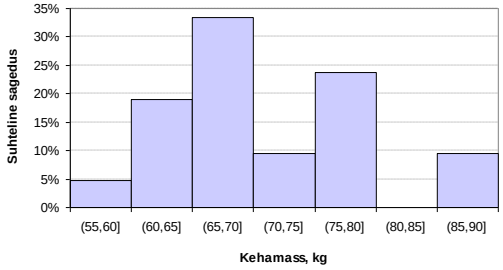
| Saagikus | Sagedus | Osakaal | Osakaal (%) | Jaotus |
|---|---|---|---|---|
| [3500-4000) | 2 | 0,2 | 20% | 20% |
| [4000-4500) | 3 | 0,3 | 30% | 50% |
| [4500-5000) | 4 | 0,4 | 40% | 90% |
| [5000-5500) | 1 | 0,1 | 10% | 100% |




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

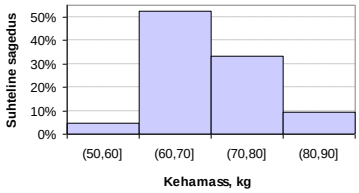
## Veel märkusi ja soovitusi

Erinevalt tulpdiagrammist, mis on antud andmete korral üheselt määratud, võime samade andmete põhjal saada üsna erineva kujuga histogramme.



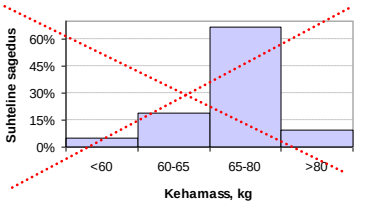

| Lammaste kehamass, kg | Sagedus | Suhteline sagedus |
|---|---|---|
| (55,60] | 1 | 4,76% |
| (60,65] | 4 | 19,05% |
| (65,70] | 7 | 33,33% |
| (70,75] | 2 | 9,52% |
| (75,80] | 5 | 23,81% |
| (80,85] | 0 | 0,00% |
| (85,90] | 2 | 9,52% |

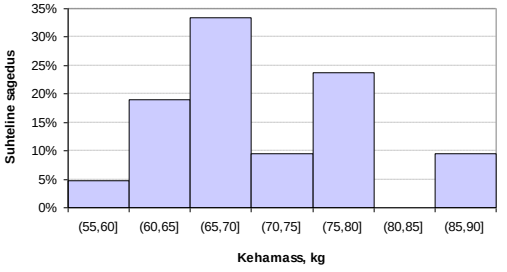
| Lammaste kehamass, kg | Sagedus | Suhteline sagedus |
|---|---|---|
| (50,60] | 1 | 4,76% |
| (60,70] | 11 | 52,38% |
| (70,80] | 7 | 33,33% |
| (80,90] | 2 | 9,52% |

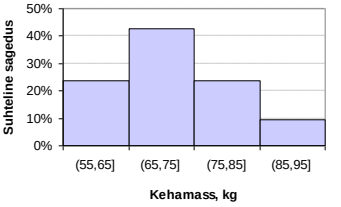


## Veel märkusi ja soovitusi

- ✓ On tungivalt soovitav, et kõik kasutatud vahemikud oleksid võrdse pikkusega!
- ✓ Joonisele tuleb kanda ka vahemikud, kuhu ühtki objekti ei sattunud!
- ✓ Avatud vahemikke tuleks võimaluse korral vältida.



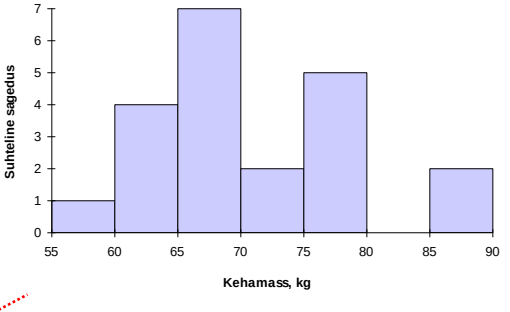



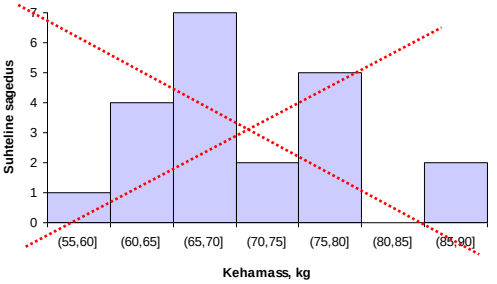




## Veel märkusi ja soovitusi

- ✓ Teaduslikult korrektsel histogrammil on ka x-telg esitatud pideval (mitte diskreetsel) skaalal.

Selle *Excel*'is teostamine on muidugi üks paras nikerdus ...
(http://ph.emu.ee/~ktanel/joonised_excelis/joonis3.php)



## Tegelikkus (mõnikord)



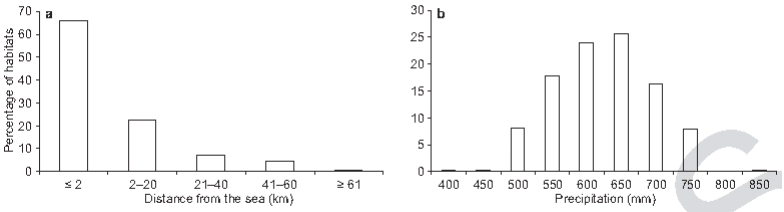
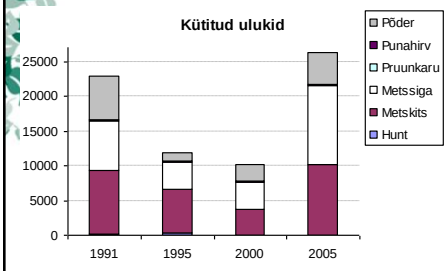
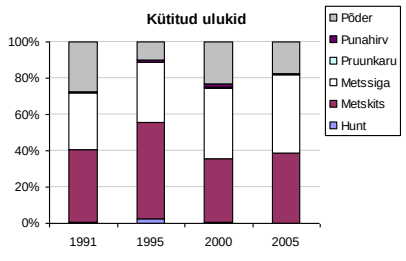



**Fig. 3.** The habitats of *Peniophora junipericola*. — **a**: Distance of the localities from the sea; — **b**: Localities partitioned by amount of precipitation.

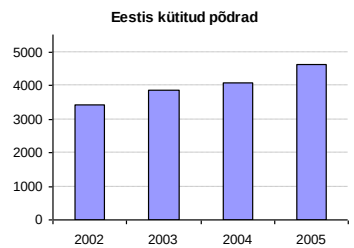
## Veel märkusi ja soovitusi

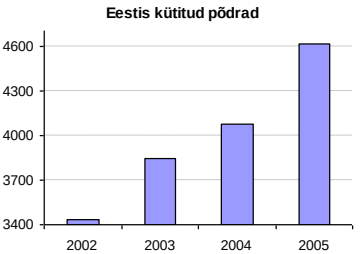
✓ Sagedused *versus* osakaalud




✓ Skaala!



✓ Vahel võib kaaluda absoluutsete ja suhteliste sageduste esitamist samal joonisel ...

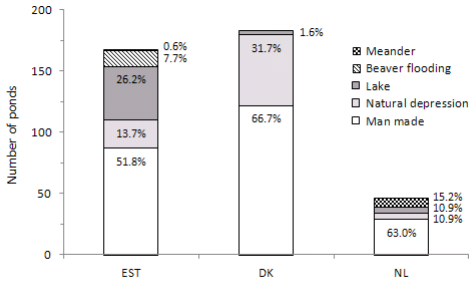
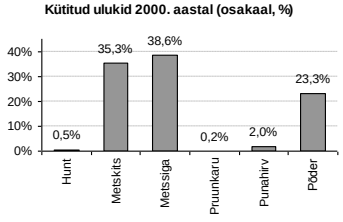
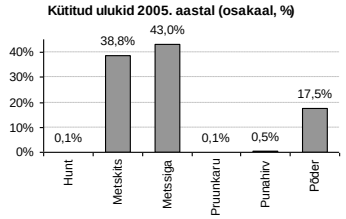


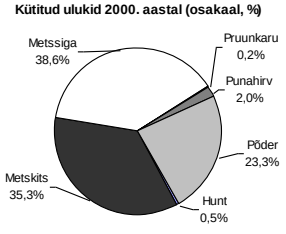

Figure 2. Amount and percentage of studied ponds depending on the pond's type.

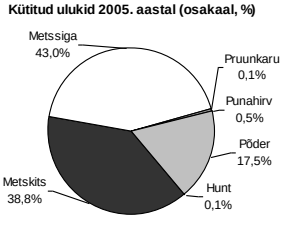
## Veel märkusi ja soovitusi

✓ Ringdiagrammile eelistada tulpdiagrammi (eriti võrdlemisel).



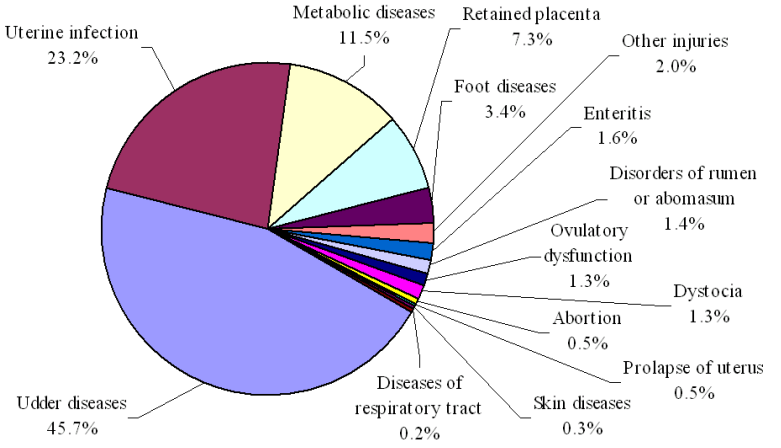



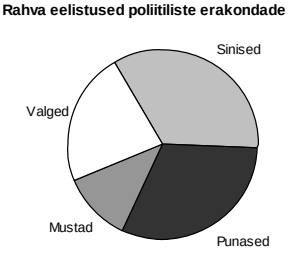
✓ Sektordiagrammi kasutage üksnes juhul, kui te tõesti soovite kirjeldada, kuidas tervik osadeks jaguneb (ja mitte seda, millist väärtust rohkem, millist vähem esineb).

Figure 1. The panorama of multifactorial diseases in Estonian dairy cows

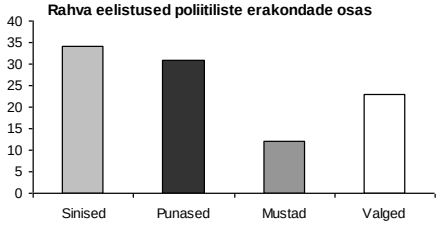


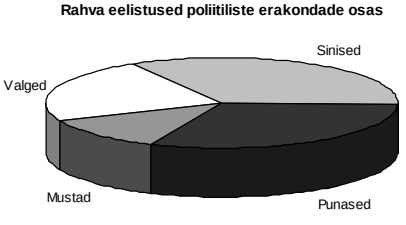

## Veel märkusi ja soovitusi

✓ Vältida tuleks 3-mõõtmelisi graafikuid, eriti ringdiagramme.




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



## Arvkarakteristikud

- andmestiku suurus (valimi maht, *sample size*) – $n$
- (aritmeetiline) keskmine [*average, mean*] – $\bar{x} = \sum_{i=1}^{n} x_i \big/ n$
- 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*

## Keskmise omadusi

1. $\overline{cx} = c\bar{x}$, kus $c$ on konstant
2. $\overline{x+c} = \bar{x} + c$, kus $c$ on konstant
3. $\overline{x+y} = \bar{x} + \bar{y}$
4. $\sum_{i=1}^{n} x_i = n\bar{x}$
5. $\overline{f(x)} \neq f(\bar{x})$, kus $f$ on monotoonne teisendus

## Mediaani omadusi

1. $med\ f(x) = f\ med(x)$, kus $f$ on monotoonne teisendus
   Näiteks, kui $med\ \log_{10}(x) = 2$,
   siis $\log_{10}\ med(x) = 2 \Rightarrow med(x) = 10^2 = 100.$
2. $\sum_{i=1}^{n} x_i \neq n \times med(x)$

## Vaatluste hajuvus

- miinimum, maksimum, haare [*range*] = max – min
- standardhälve [*standard deviation*] – $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- dispersioon [(*sample*) *variance*] – $s^2$
- standardviga [*standard error*] – $se = s/\sqrt{n}$

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

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

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

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

## Standardhälbe ja dispersiooni omadusi

1. $s^2(cx) = c^2 s^2(x)$, kus $c$ on konstant
2. $s(cx) = cs(x)$
3. $s^2(x+c) = s^2(x)$
4. $s(x+c) = s(x)$
5. kui $x$ ja $y$ on sõltumatud uuritavad tunnused, siis
   $s^2(x+y) = s^2(x) + s^2(y)$

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 asuvad keskväärtusele lähemal kui kolm standardhälvet (enamasti asub kolme standardhälbe kaugusel keskväärtusest rohkem kui 99% vaatlustest).

# Arvkarakteristikud

**Näiteid kirjandusest**

**A comparison of the methods for determination of the rennet coagulation properties of milk**

*Acta Agriculturae Scand Section A*, 2005; 55: 145–148

Taylor & Francis
Taylor & Francis Group

IVI KÜBARSEPP[1], MERIKE HENNO[1], OLAV KÄRT[1] & TUOMO TUPASELA[2]

[1]*Department of Animal Nutrition, Institute of Veterinary Medicine and Animal Science, Estonian University of Life Sciences, Kreutzwaldi 48, 51006 Tartu, Estonia, and* [2]*MTT Food Research, Myllytie 1, 31600, Jokioinen, Finland*

Table I. Means, ranges and standard deviations (SD) for milk composition and rennet coagulation parameters.

| | Mean | Min. | Max. | SD |
|---|---|---|---|---|
| Fat,% | 3.94 | 2.70 | 8.08 | 0.790 |
| Protein,% | 3.41 | 2.56 | 4.62 | 0.456 |
| Lactose,% | 4.81 | 4.38 | 5.18 | 0.174 |
| Formagraph | | | | |
| RCT, min | 9.5 | 3.5 | 35 | 4.95 |
| $E_{30}$, mm | 26.3 | 0 | 52 | 10.34 |
| Optigraph | | | | |
| $R_{initial}$, min | 6.63 | 3.73 | 19.00 | 2.423 |
| R, min | 9.53 | 4.36 | 31.59 | 4.320 |
| $A_{30}$, V | 13.72 | 0 | 35.98 | 5.855 |

# Arvkarakteristikud

**Näiteid kirjandusest**

ISSN 1392-2130. VETERINARIJA IR ZOOTECHNIKA. T. 36 (58). 2006

EFFECT OF BODY CONDITION SCORE AT PARTURITION ON THE PRODUCTION PERFORMANCE, FERTILITY AND CULLING IN PRIMIPAROUS ESTONIAN HOLSTEIN COWS

Jaak Samarütel, Katri Ling, Hanno Jaakson, Tanel Kaart, Olav Kärt
*Institute of Veterinary Medicine and Animal Sciences, Estonian University of Life Sciences, 1 Kreutzwaldi St, 51006, Tartu, Estonia, tel. +372-7-313-476, fax: +372-7-313-477; e-mail: Jaak.Samarytel@emu.ee*

Table 2. **Fertility parameters of the first parity Estonian Holstein cows grouped by BCS at parturition**

| Fertility parameters | Body condition score at calving | | |
|---|---|---|---|
| | ≤3.0 (n = 26) | 3.25–3.5 (n = 39) | ≥3.75 (n = 21) |
| Interval calving to first service (days) | 91 ± 4.1 | 83 ± 3.5 | 88 ± 5.4 |
| First service conception rate (%) | 17 | 23 | 0 |
| Service period (days) | 82 ± 14.4 | 72 ± 13.9 | 77 ± 15.8 |
| Days open (of those pregnant) | 173 ± 13.7 | 155 ± 14.8 | 165 ± 16.6 |
| Services per conception | 3.0 ± 0.36 | 3.0 ± 0.32 | 3.6 ± 0.42 |
| Number of cows not pregnant | 1 | 5 | 5 |

Values are arithmetical means ± S.E.

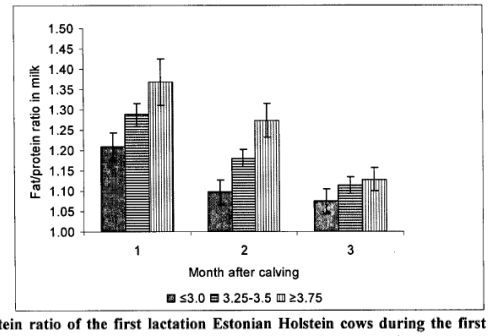



Figure 2 **Milk fat/protein ratio of the first lactation Estonian Holstein cows during the first 3 months after calving.** Values are means ± S.E. Milk fat/protein ratio was different ($P<0.05$) between BCS ≤3.0 (*thin*) and ≥3.75 (*fat*) groups during the first and second months of lactation

- ✓ Tabelite puhul ei kasutata (enamasti) vertikaalseid jooni, selle asemel „mängitakse“ horisontaalsete joontega (ja vajadusel ka tühjade veergudega).

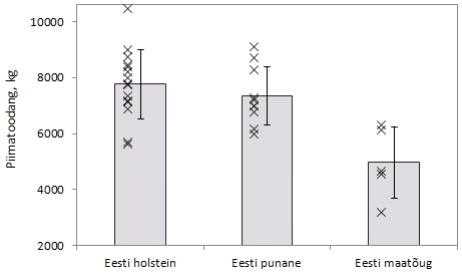
Table 2. Mean pod and hay yields (t ha$^{-1}$) of five groundnut cultivars grown at Wad Medani, Sudan in 1984–86.

| | Pod yield | | | | Hay yield | | | |
|---|---|---|---|---|---|---|---|---|
| Cultivar | 1984 | 1985 | 1986 | Mean | 1984 | 1985 | 1986 | Mean |
| Early Bunch | 2.47 | 3.34 | 2.14 | 2.65 | 3.56 | 2.40 | 2.61 | 2.86 |
| Georgia 119–20 | 1.73 | 3.38 | 2.29 | 2.46 | 3.50 | 3.88 | 2.81 | 3.39 |
| UF79–1499 | 1.61 | 2.05 | 2.34 | 2.00 | 5.23 | 1.82 | 3.26 | 3.44 |
| Apollo | 1.47 | 3.38 | 1.93 | 2.26 | 4.65 | 3.44 | 2.81 | 3.63 |
| MH 383 | 1.75 | 3.70 | 2.14 | 2.53 | 4.78 | 4.58 | 2.99 | 4.12 |
| Mean | 1.81 | 3.17 | 2.17 | 2.38 | 4.34 | 3.22 | 2.88 | 3.49 |
| s.e. | 0.235 | 0.223 | 0.167 | 0.210 | 0.571 | 0.292 | 0.264 | 0.373 |

| Line | Number of stallions | Inside lines | | | | Between lines | |
|---|---|---|---|---|---|---|---|
| | | Coancesty coefficient | | Inbreeding coefficient | | Coancesty coefficient | |
| | | Average | Max | Average | Max | Average | Max |
| Ahti | 7 | 0,092 | 0,276 | 0,012 | 0,029 | 0,037 | 0,161 |
| Eni | 4 | 0,110 | 0,255 | 0,004 | 0,009 | 0,041 | 0,163 |
| Raspel | 9 | 0,090 | 0,287 | 0,014 | 0,047 | 0,040 | 0,163 |
| Taru | 1 | 0,000 | 0,000 | 0,002 | 0,002 | 0,026 | 0,088 |
| Taube | 2 | 0,074 | 0,074 | 0,012 | 0,020 | 0,028 | 0,088 |

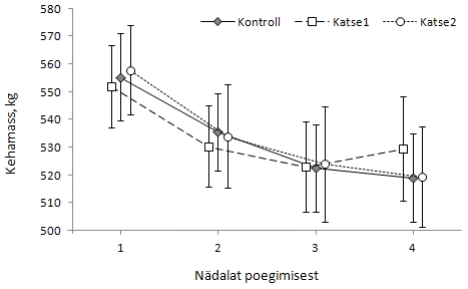
| Line | Number of stallions | Inside lines | | | | Between lines | |
|---|---|---|---|---|---|---|---|
| | | Coancesty coefficient | | Inbreeding coefficient | | Coancesty coefficient | |
| | | Average | Max | Average | Max | Average | Max |
| Ahti | 7 | 0,092 | 0,276 | 0,012 | 0,029 | 0,037 | 0,161 |
| Eni | 4 | 0,110 | 0,255 | 0,004 | 0,009 | 0,041 | 0,163 |
| Raspel | 9 | 0,090 | 0,287 | 0,014 | 0,047 | 0,040 | 0,163 |
| Taru | 1 | 0,000 | 0,000 | 0,002 | 0,002 | 0,026 | 0,088 |
| Taube | 2 | 0,074 | 0,074 | 0,012 | 0,020 | 0,028 | 0,088 |

## Veel näiteid

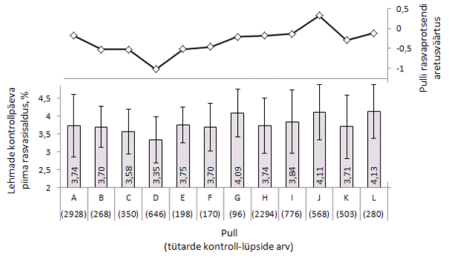



✓ Väikese andmestiku korral võib algandmed esitada samal joonisel keskmistega ...

✓ Ajas järgnevate vaatluste korral võib erinevate gruppide keskmisi veajoonte kattumise vältimiseks pisut nihutada …




✓ Vahel võib ühendada mitu joonist üheks jooniseks ...




**Joonis 20.** Lehmade esimese laktatsiooni kontroll-lüpside keskmine (± standardhälve) piima rasvasisaldus pullide kaupa ja pullide rasvaprotsendi aretusväärtus (Jõudluskontrolli Keskuse andmeil seisuga 01.01.2011)

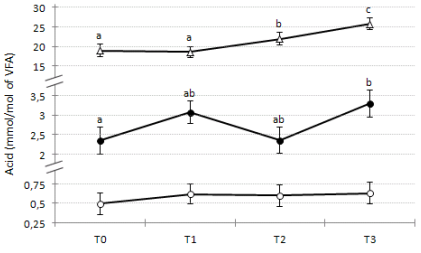


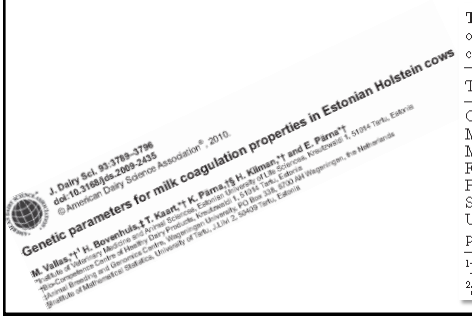

**Fig. 1.** Least square means (± standard error) of isobutyric (○),valeric (●) and propionic (Δ) acid depending on the diet. $^{abc}$ Least square means with different letters are statistically significantly different ($P < 0{\cdot}05$).

• **Variatsioonikordaja** [*coefficient of variation*] – $v = s/\bar{x} \times 100\%$

Aga mis siis, kui keskmine on negatiivne?

Näide.

| | Piim, kg | Rasv, % | Valk, % | SRA, tuh/ml | Energia-bilanss, MJ |
|---|---|---|---|---|---|
| Keskmine | 30,23 | 4,13 | 3,17 | 695,92 | -36,24 |
| St. hälve | 5,32 | 0,74 | 0,24 | 1111,99 | 52,99 |
| Var. kordaja | 17,60 | 17,98 | 7,59 | 159,79 | -146,22 |


J. Dairy Sci. 93:3793–3796
doi:10.3168/jds.2009-2435
© American Dairy Science Association®, 2010.

Genetic parameters for milk coagulation properties in Estonian Holstein cows

M. Vallas,*† H. Bovenhuis,‡ T. Kuurt,*† K. Pärna,*§ H. Kilman,*† and E. Pärna*†

**Table 1.** Number of observations (n), means, and coefficients of variation (CV) for test-day milk coagulation, production, and composition traits

| Trait | n | Mean | CV (%) |
|---|---|---|---|
| Curd firmness ($E_{30}$) (mm) | 17,577 | 27.0 | 27 |
| Milk coagulation time (RCT)[1] (min) | 17,577 | 2.3 | 9 |
| Milk yield (kg) | 17,575 | 25.9 | 28 |
| Fat (%) | 17,536 | 4.05 | 17 |
| Protein (%) | 17,567 | 3.38 | 9 |
| SCS[2] | 17,567 | 2.9 | 65 |
| Urea (mg/L) | 17,223 | 26.8 | 31 |
| pH | 17,577 | 6.6 | 1 |

[1]Log-transformed.
[2]SCS = [$\log_2$(SCC/100,000) + 3].

# Kvantiilid, protsentiilid

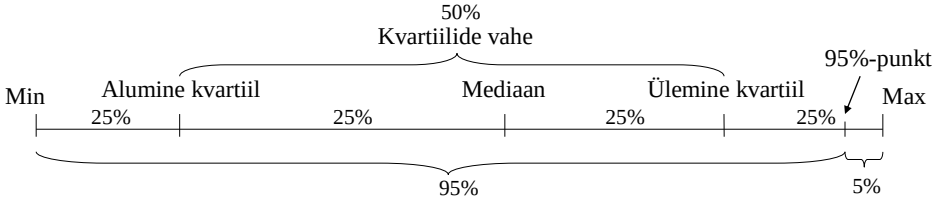
• **kvartiilid** – alumine kvartiil e 25%-punkt ja ülemine kvartiil e 75%-punkt [*lower*, *upper quartile*]

• **kvartiilide vahe** [*interquartile range*, *IQR*] – kasutatakse varieeruvuse iseloomustamiseks

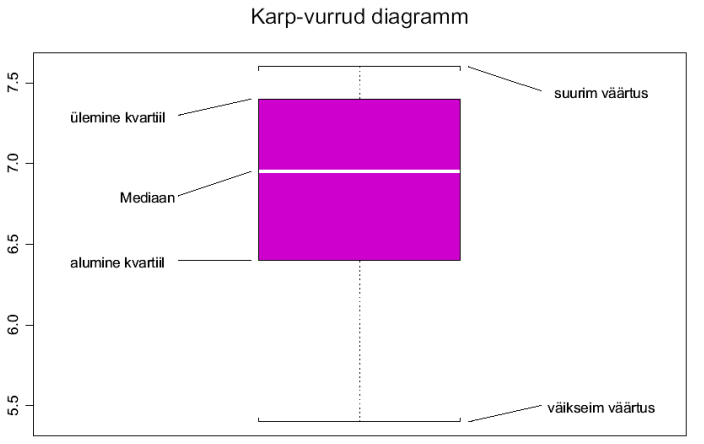
• detsiilid, protsentiilid e protsendipunktid/**kvantiilid**

> $\alpha$-kvantiiliks [*$\alpha$-quantile*] nimetatakse sellist uuritava tunnuse väärtust, millest väiksemate väärtuste osakaal mõõtmistulemuste seas on $\alpha$.

• min, max



# Karp-vurrud diagramm [*boxplot*]




# Karpvurrud-diagramm

**Näiteid kirjandusest**

Coping capacity of dairy cows during the change from conventional to automatic milking[1]

D. Weiss*, S. Helmreich*, E. Möstl†, A. Dzidic*, and R. M. Bruckmaier*[2]

*Physiology Weihenstephan, Technical University, Munich, Germany and
†Institute of Biochemistry, University of Veterinary Medicine, Vienna, Austria

 J. Anim. Sci. 2004. 82:563–570

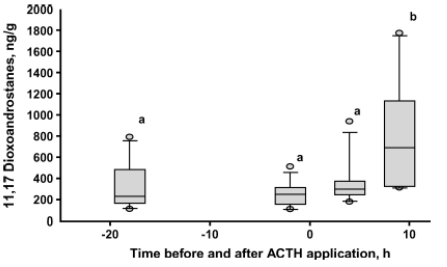
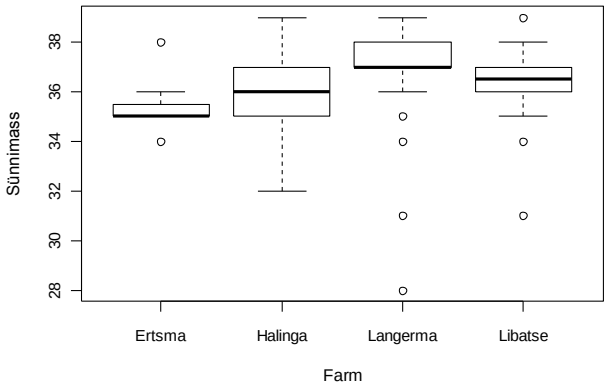


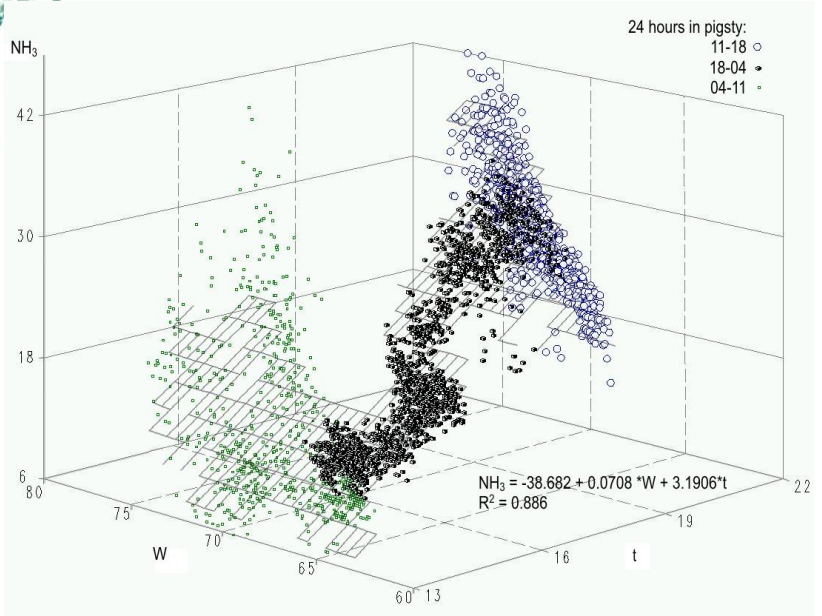

**Figure 9.** Boxplot of the feces 11,17 dioxoandrostanes concentrations (fresh-matter basis) before and during ACTH administration. Each box shows the median and the upper and lower quartile value; the whiskers show the 10th and the 90th percentiles. The circles represent data points that were outside the centiles. Means without common letters differ ($P < 0.05$).

## Karpvurrud-diagramm

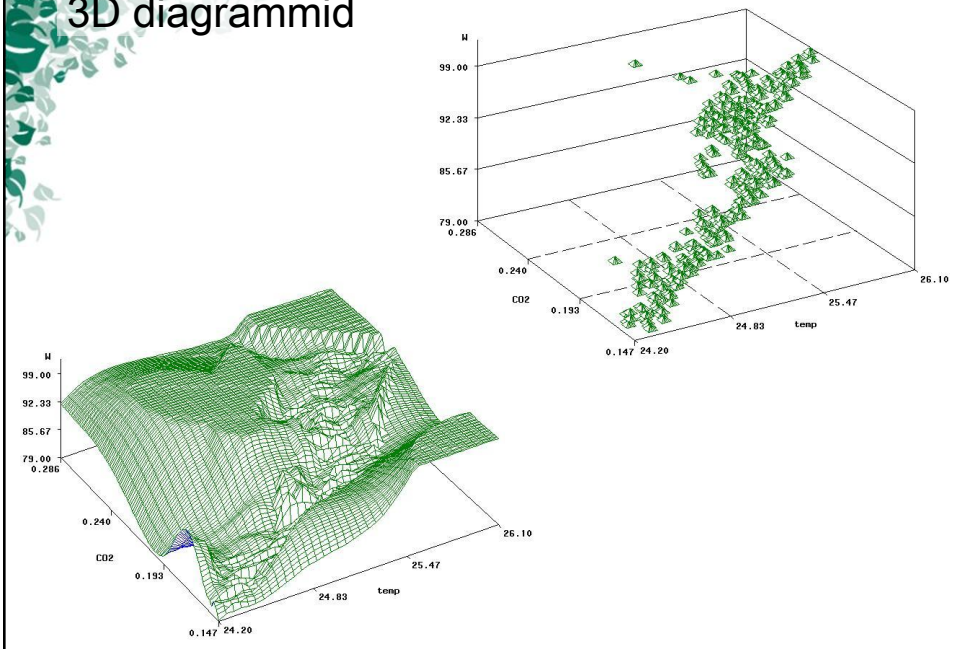



Vasikate sünnimass farmide kaupa. Väärtused, mis jäävad alumisest või ülemisest kvartiilist kaugemale kui 1,5-kordne kvartiilide vahe, on loetud erandlikeks ja tähistatud sümboliga °.

## 3D diagrammid



## 3D diagrammid




✓ 3-mõõtmelised pinnad sobivad sageli illustreerimaks matemaatilise modelleerimise tulemusi.

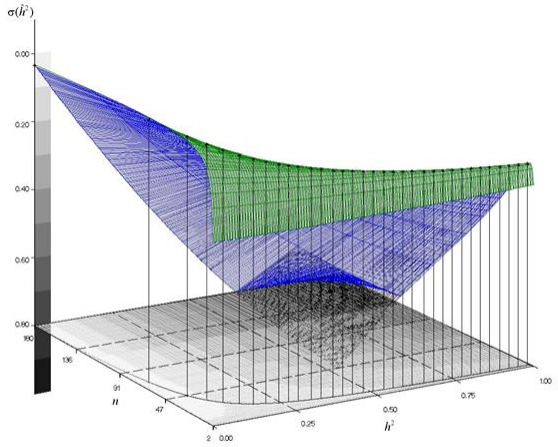
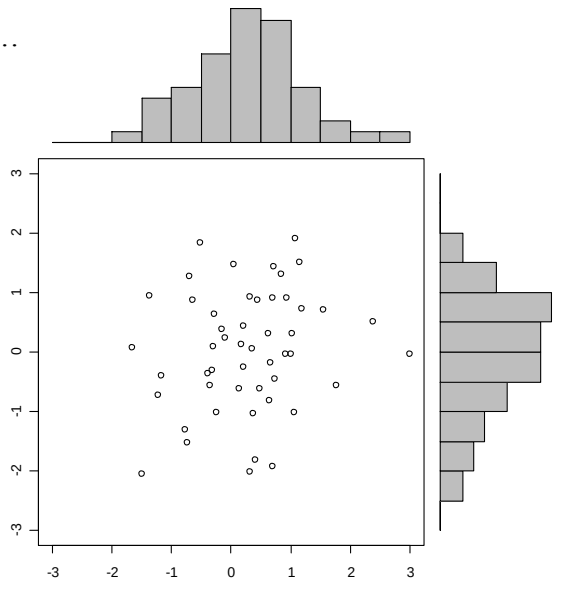
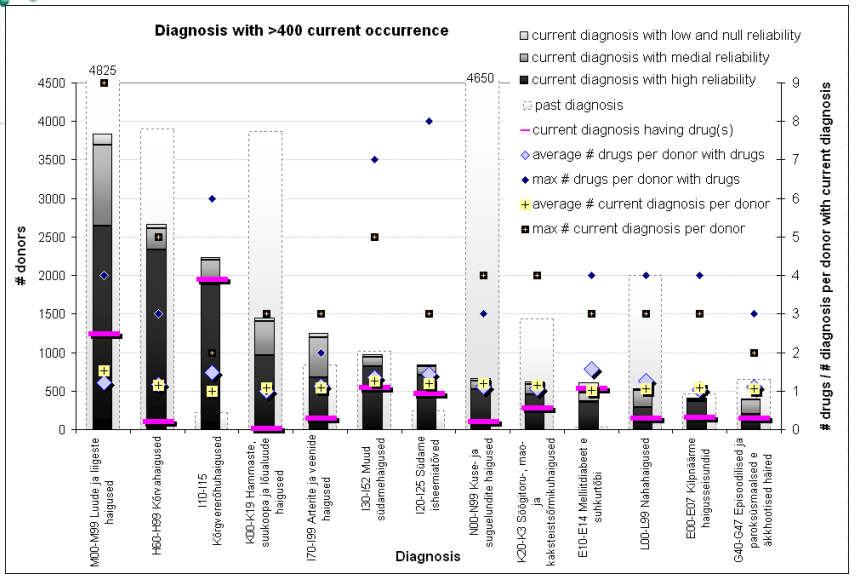



**Figure 3.4.** Pattern of $\sigma(\hat{h}^2)$ and optimal number of daughters per sire (vertical arrows for integer numbers and dotted line on $xy$-plane for continuous numbers) in different true heritability values ( $N = 360$ , $\sigma_e^2 = 1$ ).

✓ Diagramme võib omavahel kombineerida …




✓ Mõnikord võib terve ettekande mahutada ühele joonisele ...



# Populatsioon ja valim

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

## Populatsioon *versus* valim

**Üldkogum** (**populatsioon**) on reaalselt olemasolev või ka abstraheeritud objektihulk, mille/kelle kohta soovitakse uurimistöö tulemusena sisulisi järeldusi teha
Populatsiooni defineerides piiritletakse ära uuritav objekt ajas, ruumis, katsetingimuste kaudu, …

Näiteks:
- Eesti talud 2007. aastal,
- põllu nr 12154 saagikus (nii minevikus, praegu kui ka tulevikus),
- kõik Eestimaa põllud, millel kasvatatakse talirukist,
- Eesti põhjaranniku jõgedes kudevad lõhed,
- Eesti maatõugu veised,
- Eesti sigade pekipaksuse muutus ajavahemikul 1995-2005,
- kõik antud mündiga teha võidavad kulli/kirja viskamised,
- …

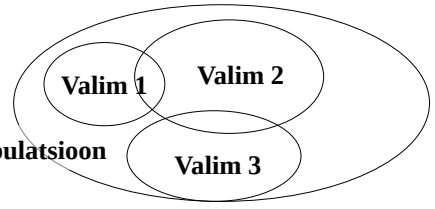
# Populatsioon *versus* valim

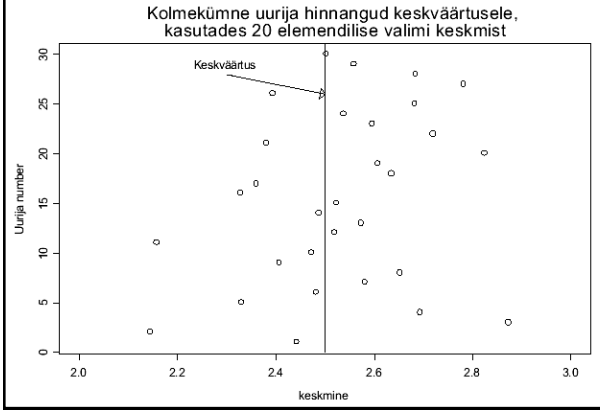
**Valim** [*sample*] on teatava eeskirja järgi moodustatud hulk üldkogumisse kuuluvatest objektidest e uurimiseks valitud üldkogumi osa.

| Populatsioon | Valim |
|---|---|
| Eesti vetes kudevad lõhed | Kontrollpüükidel püütud 60 lõhet |
| Eestis jõudluskontrolli all olevad 1. laktatsiooni EHF-lehmad | Jõudluskontrollikeskuse andmebaasist välja valitud 12000 1. laktatsiooni EHF-lehma |
| Kolm erineva säilitusainega jogurtipartiid | 10 proovi igast jogurtipartiist |

Statistika olemus:
- võtame teatud reeglite järgi osa üldkogumist (valimi),
- analüüsime seda ja
- teeme järeldusi kogu üldkogumi kohta!

# Populatsiooni parameetrite hindamine



## Punkthinnangud

Kui huvipakkuva väärtuse hinnanguks on üks konkreetne arv, siis räägitakse, et tegemist on **punkthinnanguga**.
Märkimaks, et tegu on hinnanguga, kirjutatakse sageli arvkarakteristikut iseloomustava sümboli kohale katuseke, laineke vms; näiteks parameetri $\theta$ hinnang on $\hat{\theta}$.

| Populatsiooni parameeter | Hinnang valimi põhjal |
|---|---|
| keskväärtus [*expectation*]<br>E$X$, $\mu$ | keskmine<br>$\bar{x}$ |
| populatsiooni dispersioon<br>D$X$, var($X$), $\sigma^2$ | valimi dispersioon<br>$\hat{\mathrm{D}}X$, v$\hat{\mathrm{a}}$r($X$), $s^2$ |
| populatsiooni standardhälve<br>$\sigma$ | valimi standardhälve<br>$s$ |
| populatsiooni mediaan<br>med($X$) | valimi mediaan<br>m$\hat{\mathrm{e}}$d($X$) |
| populatsiooni $\alpha$-kvantiil<br>$q_\alpha$ | valimi $\alpha$-kvantiil<br>$\hat{q}_\alpha$ |

## Populatsioon *versus* valim

Valimi põhjal tehtud järeldused on õiged,

- kui valim on **esindav** e representatiivne, st et uuritava tunnuse väärtuste proportsionaalne jaotus valimis on enam-vähem samasugune kui populatsioonis ning
- kui otsuste tegemisel on järgitud matemaatilise statistika reegleid.

Andmete kogumisel tuleb

- vältida teadlikult üksnes soovitud tendentsi peegeldavate vaatluste registreerimist;
- püüda tagada kõigi uuritavate objektide "võrdne kohtlemine" uuringus mittehuvipakkuvate näitajate osas; kui viimane pole võimalik, tuleks need nn segavad faktorid samuti registreerida võimaldamaks hilisemal analüüsil nende (potentsiaalse) mõju arvesse võtmist.

Kui valim ja populatsioon kattuvad, on tegu **kõikse uuringuga**.
NB! Oma loomult välistab kõikne uuring igasugused teaduslikel alustel tehtavad prognoosid!

# Tõenäosus ja teoreetilised jaotused

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

Statistilise analüüsi tulemused ei ole kunagi 100% kindlad!

Sestap taandub küsimus pigem sellele, kui õiged või valed, täpsed või ebatäpsed saadud tulemused on.

Viimase hindamiseks on vaja tõenäosuse mõistet, mingit täpsuse mõõtu ning eeskirju nende leidmiseks.

**Tõenäosus** [*probability*] on sündmuse toimumise mõõt skaalal 0-st 1-ni.

Võimatu sündmuse tõenäosus on 0 ja kindla sündmuse tõenäosus on 1.

Tõlgendused:

tõenäosus – osakaal e protsent
(kui teatud haigus esineb 5%-l uuritavast populatsioonist, siis tõenäosus, et selle populatsiooni juhuslikult valitud esindaja on haige, on 0,05 ehk 5%);

tõenäosus – usk sündmuse või nähtuse võimalikkusesse.

## Tõenäosus

**Kuidas leida praktikas sündmuse toimumise tõenäosust?**

• Teha väga palju (teoreetiliselt lõpmatu arv) katseid/vaatlusi – nn statistiline tõenäosus – sündmuse toimumise suhteline sagedus annab tulemuseks ligikaudse tõenäosuse.

• Tundes uuritavat objekti hästi või mõistes täielikult, mis katse käigus toimub, on võimalik tõenäosust leida mõttetöö tulemusena (st, et teatavate sündmuste toimumise tõenäosuse kohta saab teha äärmiselt usutavaid eeldusi) – nn klassikaline tõenäosus (see, mida koolis õpitakse).

• Vahel on võimalik sündmuse tõenäosust arvutada, kui teame mõne teise sündmuse toimumise tõenäosust.

## Statistiline tõenäosus

**Suurte arvude seadus**: katseseeria lõpmatul pikenemisel läheneb sündmuse suhteline sagedus tema tõenäosusele.

Suhtelise sageduse kaudu leitud nn statistiline tõenäosus on teoreetilise tõenäosuse hinnanguks (st, et ei ole konstant – muutub katseseeria pikenedes).

Näiteks veeretate 75 korda täringut ja saate 52 korda 6 silma. Antud täringuga 6 silma saamise tõenäosus on siis hinnatav suhtest

$$\hat{P}(6\text{ silma}) = 52/75 \approx 0,693.$$

## Klassikaline tõenäosus

**Juhuslik katse** – katse, mille tulemus pole ette teada.
**Juhuslik sündmus** – juhusliku katse tulemus.

$$\text{Tõenäosus} = \frac{\text{Sündmuse jaoks soodsate katsetulemuste arv}}{\text{Kõigi katsetulemuste arv}}$$

Näide. Katseks on 20-tahulise täringu veeretamine, sündmuseks $A$ on 10-ga jaguva silmade arvu saamine.

$$\mathrm{P}(A) = 2 / 20 = 0{,}1 .$$

## Tehted tõenäosustega

1. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
2. Tinglik tõenäosus $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) / \mathrm{P}(B)$,
   millest $\mathrm{P}(A \cap B) = \mathrm{P}(B) \times \mathrm{P}(A \mid B) = \mathrm{P}(A) \times \mathrm{P}(B \mid A)$
3. Kui $A$ ja $B$ on üksteist välistavad, siis
   $\mathrm{P}(A \cap B) = 0$,
   $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$
4. Sõltumatute sündmuste korral
   $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$,
   $\mathrm{P}(A \mid B) = \mathrm{P}(A)$ ja $\mathrm{P}(B \mid A) = \mathrm{P}(B)$
5. Täistõenäosuse valem: $\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(H_i) \times \mathrm{P}(A \mid H_i)$
6. Bayesi valem: $\mathrm{P}(H_i \mid A) = \dfrac{\mathrm{P}(A \mid H_i) \times \mathrm{P}(H_i)}{\sum_{j=1}^{k} \mathrm{P}(A \mid H_j) \times \mathrm{P}(H_j)}$

## Teoreetilised jaotused

• Teoreetilised jaotused leitakse analüütiliselt või arvutisimulatsioonide abil, **baseeruvana uuritava tunnuse** (=juhusliku suuruse) või selle väärtuste funktsiooni (e statistiku) tekkemehhanismil e **olemusel**.

Näiteks on matemaatika mõistes oma olemuselt sarnased tunnused 'bakterite arv 1 ml piimas', 'emise pesakonna suurus', 'edukalt talvitunud mesitarude arv mesilas' jne, või siis 'lõhe kasvukiirus', 'õhu liikumise kiirus laudas', 'mulla happesus' jne.

• **Teoreetilised jaotused kirjeldatakse parameetritest sõltuvate eeskirjadega**, mille abil on võimalik leida vastava jaotusega tunnuste (statistikute) väärtuste esinemise tõenäosused.

• **Teoreetilised jaotused on aluseks teaduslike järelduste tegemisel** (statistiliste hüpoteeside kontrollimisel, sageli ka parameetrite väärtuste hindamisel ja nende hinnangute usaldusväärsuse leidmisel).

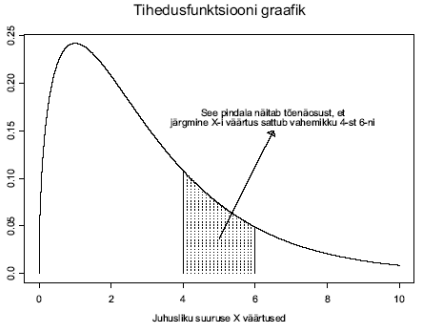
Seejuures on järeldused õiged üksnes siis, kui nad on tehtud andmetega sobivatele teoreetilistele jaotustele tuginedes (seda eriti väikeste, $n < 100$, valimite puhul)!

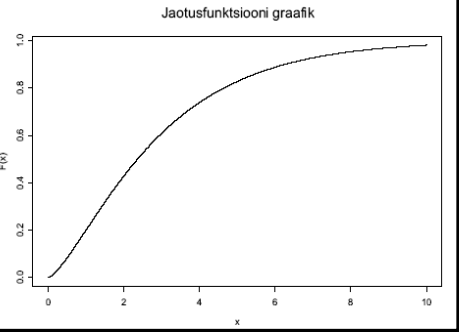
## Teoreetilised jaotused

Diskreetne jaotus esitatakse **tõenäosusfunktsiooniga $p(k) = P(X=k)$**, kus $k$ on jaotuse võimalik väärtus.

Pidev jaotus esitatakse **tihedusfunktsiooniga $f(x) = dF(x)/dx$**, kus $F(x) = P(X \leq x)$ on **jaotusfunktsiooni** väärtus kohal $x$,

$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a).$$



## Tuntumaid jaotusi

**Bernoulli jaotusega** on kõik binaarsed tunnused, $X \sim Be(p)$, kus $p$ on Bernoulli jaotuse parameeter (tõenäosus, et uuritav suurus omandab väärtuse 1).

Seejuures $E(X) = p$ ja $D(X) = p\ (1 - p)$.

**Binoomjaotus**

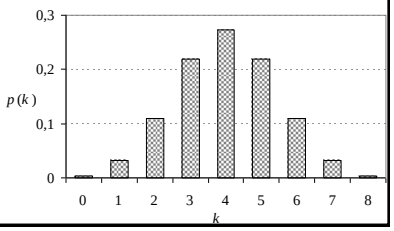
Sündmuse toimumiste arv $n$-katselises katseseerias, kus igal üksikul katsel on sündmuse toimumise tõenäosus $p$: $X \sim B(n;p)$.

Tõenäosusfunktsioon: $p(k) = C_n^k p^k (1-p)^{n-k}$, $C_n^k = \frac{n!}{k!(n-k)!}$

$E(X) = np$ ja $D(X) = np(1 - p)$

Näited. Koeral sündis 8 kutsikat. Huvi pakkuv suurus $X$ on isaste arv nende hulgas. Lihtsuse mõttes on eeldatud, et isase ja emase kutsika sündimise tõenäosus on võrdne ($p = 0{,}5$).

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $p(k)$ | 0,004 | 0,031 | 0,109 | 0,219 | 0,273 | 0,219 | 0,109 | 0,031 | 0,004 |




## Tuntumaid jaotusi

**Poissoni jaotus**

Poissoni jaotusega on näiteks ühe päeva jooksul aset leidvate südameatakkide arv Tartu linnas, raku jagunemisel tekkivate geenimutatsioonide arv jne. Seda, et tunnus $X$ on Poissoni jaotusega, tähistatakse $X \sim P(\lambda)$, kus $\lambda$ on keskmine südameatakkide arv ühes päevas või keskmine mutatsioonide arv raku jagunemisel.

Tõenäosusfunktsioon: $p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k = 0,1,...$ $E(X) = D(X) = \lambda$

**Geomeetriline jaotus**

Kui katse õnnestumise tõenäosus on $p$, siis katse number, millal katse esimest korda õnnestus, on geomeetrilise jaotusega juhuslik suurus.

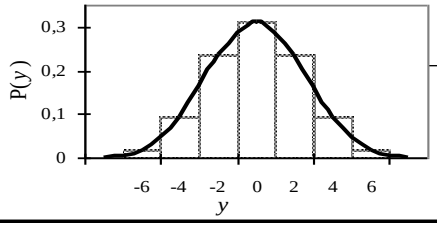
Tõenäosusfunktsioon: $p(k) = p(1-p)^{k-1}$, $k = 1,2,...$ $E(X) = (1 - p) / p$
$D(X) = (1 - p) / p^2$

## Normaaljaotus

Kui uuritavat tunnust mõjutavad paljud erinevad tegurid, millest ühegi mõju ei ole omaette võttes märkimisväärne, siis on uuritava tunnuse jaotus lähedane normaaljaotusele.

Näide. Alleelid *a*, *b*, *c* vähendavad ja *A*, *B*, *C* suurendavad fenotüübiväärtust *y* 1 võrra.

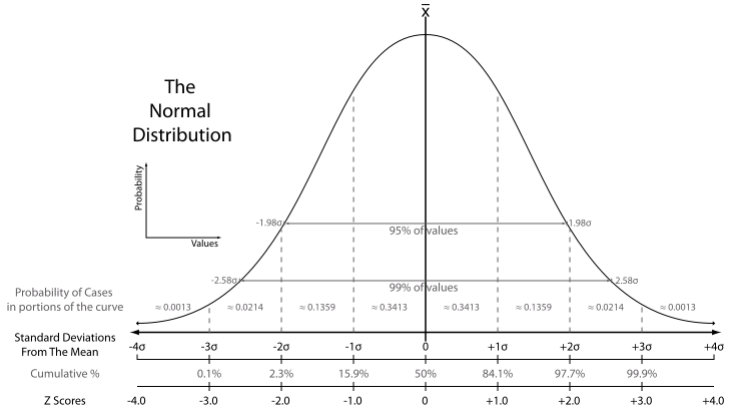
| Genotüübid (*Aa* ja *aA* jne on kirjas ühe variandina) | *y* | P(*y*) |
|---|---|---|
| AABBCC | 6 | 1/64 |
| AABBCc, AABbCC, AaBBCC | 4 | 6/64 |
| AABBcc, AABbCc, AaBBCc, AAbbCC, AaBbCC, aaBBCC | 2 | 15/64 |
| AABbcc, AaBBcc, AAbbCc, AaBbCc, aaBBCc, AabbCC, aaBbCC | 0 | 20/64 |
| AAbbcc, AaBbcc, AabbCc, aaBBcc, aaBbCc, aabbCC | -2 | 15/64 |
| Aabbcc, aaBbcc, aabbCc | -4 | 6/64 |
| aabbcc | -6 | 1/64 |




## Normaaljaotus

Tõenäosusfunktsioon: $$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

kus $\mu$ on uuritava tunnuse keskväärtus ja $\sigma^2$ dispersioon, $X \sim N(\mu,\sigma^2)$.



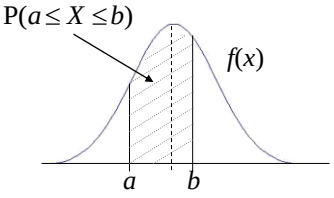
# Normaaljaotus

Normaaljaotusega juhuslike suuruste lineaarkombinatsioon on samuti normaaljaotusega (muutuvad vaid parameetrite väärtused).

Sagedasemaks lineaarteisenduseks on standardiseerimine

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1),$$

kus $N(0,1)$ on standardne normaaljaotus, mille jaotusfunktsiooni $\Phi(z) = \mathrm{P}(Z \le z)$ väärtused on tabuleeritud. Seejuures kehtivad seosed




$$F(x) = \Phi\left(\tfrac{x-\mu}{\sigma}\right) \text{ ja } \Phi(-z) = 1 - \Phi(z)$$

$$\mathrm{P}(a \le X \le b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Tabel. Standardse normaaljaotuse enamkasutatavad jaotusfunktsiooni väärtused

| $\Phi(z) = \alpha$ | 0,005 | 0,025 | 0,05 | 0,5 | 0,95 | 0,975 | 0,995 |
|---|---|---|---|---|---|---|---|
| $z_\alpha\ (q_\alpha)$ | -2,58 | -1,96 | -1,64 | 0 | 1,64 | 1,96 | 2,58 |

# Normaaljaotus

**Normaaljaotuse jaotusfunktsioon $\Phi(x)$**

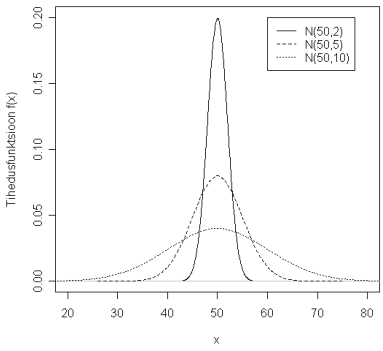
| x | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9988 | .9989 | .9989 | .9990 | .9990 |

## Normaaljaotus

**Näide.** Vere kogus indiviidi 50 ml vereproovis $X \sim N(50; \sigma^2)$, kus $\sigma$ iseloomustab proovivõtmise täpsust.

Kui suur on tõenäosus, et proovi maht erineb 50 ml-st enam kui 5 ml võrra?

$$P(X < 45 \cup X > 55) = 1 - P(45 \le X \le 55) = ?$$




$\sigma = 1$:

$$P(45 \le X \le 55) = \Phi\left(\frac{55-50}{1}\right) - \Phi\left(\frac{45-50}{1}\right)$$

$$= \Phi(5) - \Phi(-5) = \Phi(5) - [1 - \Phi(5)] = 2\Phi(5) - 1 = 2 * 0,9999997 - 1 = 0,9999994$$

$$P(X < 45 \cup X > 55) = 1 - 0,9999994 = 0,00000057 = 5,76\,E-07$$

$\sigma = 5$:

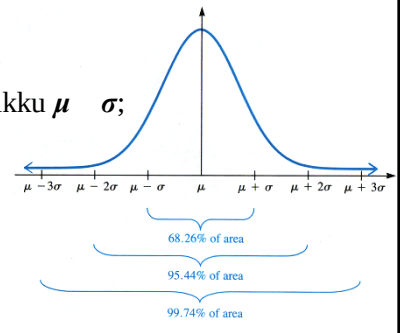
$$P(45 \le X \le 55) = \Phi\left(\frac{55-50}{5}\right) - \Phi\left(\frac{45-50}{5}\right) = 2\Phi(1) - 1 = 2 * 0,8413 - 1 = 0,6827$$

$$P(X < 45 \cup X > 55) = 1 - 0,6827 = 0,3173$$

## Normaaljaotus

Kui uuritav tunnus on normaaljaotusega, siis

- ligikaudu **68,3%** väärtustest jäävad vahemikku $\mu$ $\sigma$;
- **95,4%** vahemikku $\mu \pm 2\sigma$ ja
- **99,7%** vahemikku $\mu \pm 3\sigma$.




**Normaaljaotuse tähtsus statistikas:**

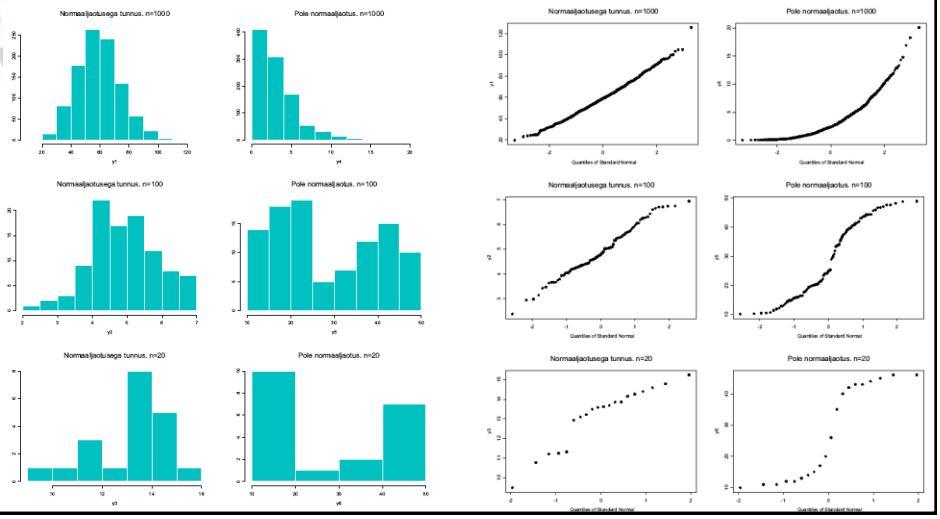
- Paljud mõõdetud tunnused on ligikaudu normaaljaotusega.
- Paljud matemaatilise statistika meetodid eeldavad tunnuse jaotumist vastavalt normaaljaotuse seaduspäradele.
- Suurte valimite korral on paljud normaaljaotusega tunnuste tarvis loodud meetoditest rakendatavad sõltumata jaotusest – näiteks läheneb mistahes jaotusega tunnuse keskmise (ja ka summa) jaotus valimi mahu kasvades normaaljaotusele.

# Normaaljaotus

**Normaaljaotuse eelduse kontrollimine graafiliselt**

Histogramm

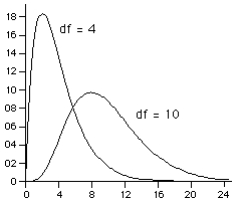
Tõenäosuspaber [*Normal prabability plot*]




# Tuntumaid jaotusi

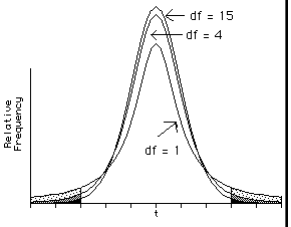
Sõltumatute standardse normaaljaotusega juhuslike suuruste $X_1,\ldots,X_n$ ruutude summa on **$\chi^2$-jaotusega** vabadusastmete arvuga $n$:

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$




Kui juhuslik suurus $X \sim N(0,1)$ ja juhuslik suurus $Y \sim \chi^2(n)$, kusjuures $X$ ja $Y$ on sõltumatud, siis $Z = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$,

st, et $Z$ on ***t*-jaotusega** vabadusastmete arvuga $n$.




Kui $X_i \sim N(\mu,\sigma^2)$, $i=1,\ldots,n$, siis

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ja $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$.

# Populatsiooni parameetrite hindamine

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

## Hinnangute leidmise meetodid

**Suurima tõepära meetod** (*maximum likelihood method*, ML-meetod)
• Suurima tõepära meetod baseerub teoreetilisel jaotusel, mille parameetriks (argumendiks) on hinnatav parameeter.
• Hinnanguks valitakse see parameetri väärtus, mis realiseerunud juhul (st uuritavate andmete korral) kõige paremini sobib ehk teisisõnu on antud valimi jaoks tõepäraseim väärtus.

**Vähimruutude meetod** (*least square method*, LS-meetod)
• Vähimruutude meetod leiab parameetri hinnangu, minimiseerides realiseerunud väärtuste (andmete) ja parameetri hinnangule vastavate väärtuste vahelise ruuterinevuse.
• Vähimruutude meetod ei eelda tihedus- või tõenäosusfunktsiooni kasutamist, mistõttu on selle abil saadavad hinnangud sageli lihtsamal kujul, võrreldes teiste hindamismeetoditega.

**Momentide meetod** (*method of moments*)

## Nihkega ja nihketa hinnangud

Parameetri $\theta$ hinnangut nimetatakse **nihketa hinnanguks** [*unbiased*], kui $\mathrm{E}(\hat{\theta}) = \theta$; ehk hinnang on „keskmiselt õige“, puudub süstemaatiline viga.

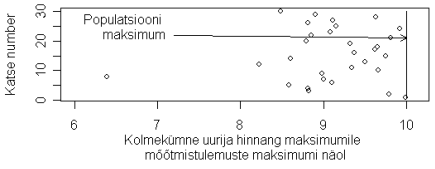

**N**äiteks valimi (juhusliku suuruse realiseerunud väärtuste) keskmine

$$\bar{x} = \tfrac{1}{n}\sum x_i$$

on nihketa hinnang populatsiooni (juhusliku suuruse) keskväärtusele $\mathrm{E}(X)$:

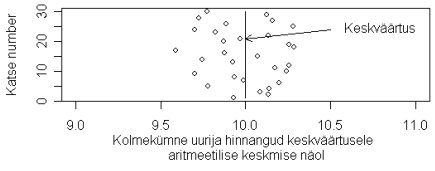
$$\mathrm{E}(\bar{x}) = \mathrm{E}(\tfrac{1}{n}\sum_{i=1}^{n} x_i) = \tfrac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{E}(x_i)}_{\mathrm{E}(X)}$$

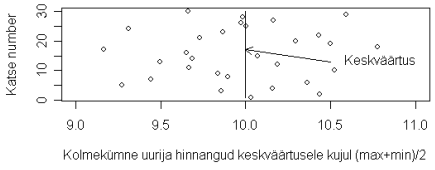
$$= \tfrac{1}{n} n\,\mathrm{E}(X) = \mathrm{E}(X).$$




## Efektiivsed hinnangud

Parameetri $\theta$ hinnangut $\hat{\theta}$ nimetatakse **efektiivseks hinnanguks**, kui $\mathrm{var}(\hat{\theta})$ on vähim kõigi parameetri $\theta$ nihketa hinnangute dispersioonide hulgas; ehk, efektiivne hinnang on täpseim hinnang.



## Hinnangu standardviga

Et andmete alusel leitud parameetri $\theta$ hinnang on juhuslik suurus, siis eksisteerib tal ka dispersioon $\text{var}(\hat{\theta})$. Viimane on aga jällegi tundmatu üldkogumi parameeter.
Seega, et saada tegelikkuses aimu oma andmete alusel leitud parameetri hinnangu täpsusest, tuleb andmetest hinnata ka hinnangu dispersioon, millest reeglina parema mõistetavuse huvides võetakse veel ruutjuur (et saada varieeruvuse hinnangut samal skaalal parameetri endaga).

Hinnangu standardhälbe hinnangut nimetatakse **hinnangu standardveaks**:

$$se(\hat{\theta}) = \sqrt{\hat{\text{var}}(\hat{\theta})}.$$

Näiteks keskväärtuse $E(X) = \mu$ hinnangu $\hat{\mu} = \bar{x}$ (valimi keskmise) dispersiooni hinnang on $\hat{\text{var}}(\hat{\mu}) = s^2/n$ ja standardviga on

$$se(\hat{\mu}) = \frac{s}{\sqrt{n}},$$

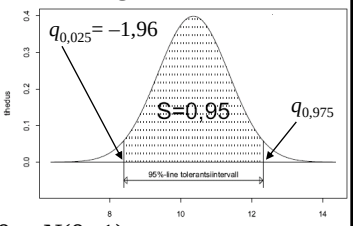
kus $s^2$ on valimi dispersioon.

## Tolerantsiintervall

Kogenud kalastajana teate, et teie poolt siiani püütud haugid on keskmiselt kaalunud 785 g standardhälbega 340 g. Perele õhtusöögiks kala püüdes oleks ju huvitav teada, kui palju kaalub järgmine konksu otsa jääv haug.

Täpset vastust sellele küsimusele statistika ei anna, küll aga saab leida vahemiku, millesse järgmise haugi kaal satub suure (näiteks 95%) tõenäosusega.

Väärtuste vahemik, kuhu kuuluvad 95% uuritava tunnuse väärtustest, on 95%-**tolerantsiintervall**.




Teame, et kui $X \sim N(785; 340)$, siis $(X–785)/340 \sim N(0; 1)$.

**Standardse normaaljaotuse kohta teame, et 95% väärtustest jääb vahemikku ($q_{0,025}$, $q_{0,975}$) = (–1,96; 1,96).**

Seega

$$\text{P}\left(-1,96 < \frac{X-785}{340} < 1,96\right) = 0,95$$

ja

$$\text{P}\ 785-1,96\times340 < X < 785+1,96\times340\ = \text{P}(118,6 < X < 1451,4) = 0,95$$

## Usaldusintervall

Vahemikhinnang (usaldusintervall, *confidence interval*, *CI*) tähendab valimi abil teatava piirkonna määramist leitud punkthinnangu ümber nii, et see piirkond kataks õige parameetri väärtuse etteantud küllalt suure tõenäosusega:

$$P(\underline{\theta} < \theta < \overline{\theta}) = 1 - \alpha,$$

• kus $1 - \alpha$ on **usaldusnivoo** [*confidence level*] (ühe lähedane, ent alati ühest väiksem);

• $\alpha$, mida nimetatakse **olulisuse nivooks** [*significance level*], on väike positiivne arv (tavaliselt 0,01 või 0,05);

• $\theta$ on õige, ent mitteteadaolev jaotusparameetri väärtus;

• $\underline{\theta}$ ja $\overline{\theta}$ on parameetri $\theta$ **(1-$\alpha$)-usalduspiirid**
(näiteks kui $\alpha = 0{,}05$, siis on tegu 95%-liste usalduspiiridega).

Täpsuse huvides räägitakse vahel ka alumisest ja ülemisest usalduspiirist [*lower*/*upper confidence limit*].

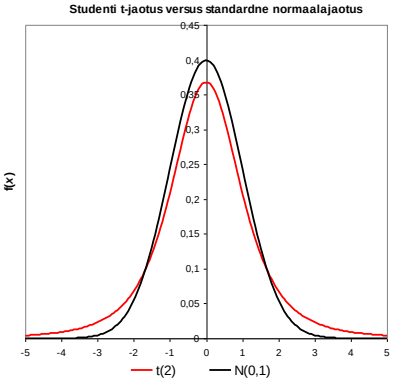
## Usaldusintervall

**Usaldusintervall** (*confidence interval*) **keskmisele**

✓ $X \sim N(\mu, \sigma^2) \Rightarrow \overline{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ ehk $\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Kui $\sigma$ ei ole teada, siis $\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$

Toodud seosed kehtivad <u>suure valimi</u> korral sõltumata uuritava tunnuse jaotusest!




✓ $\alpha$-kvantiil $q_\alpha$ ($t$-jaotuse puhul $t_{\alpha,n\text{-}1}$, standardse normaaljaotuse puhul $z_\alpha$)

$$P(X < q_\alpha) = \alpha$$

Sisuliselt sama, mis protsentiil;
näiteks 0,5-kvantiil on mediaan, sest $P(X < med) = 0{,}5$.

# Usaldusintervall

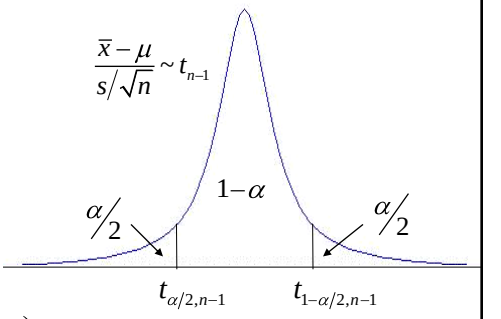
**Usaldusintervall keskmisele**

$$\mathrm{P}\left(\frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right) = \alpha/2 \quad \text{ja}$$

$$\mathrm{P}\left(\frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{1-\alpha/2,n-1}\right) = 1-\alpha/2$$

$$\Downarrow$$

$$\mathrm{P}\left(\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

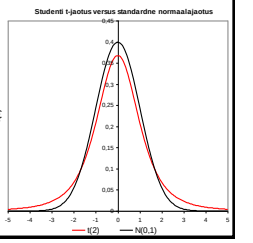



$(1-\alpha)$-usalduspiirid:

$$\underline{\mu};\bar{\mu} = \left(\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

Kui valim on suur ($n > 100$), siis võib kasutada ka normaaljaotust:

$$\underline{\mu};\bar{\mu} = \left(\bar{x} - q_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + q_{1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$




# Usaldusintervall

**Näide.** Kanauurijat Hans Hane huvitab, mitu muna munevad keskmiselt Eestis peetavad sassexi tõugu kanad ühe nädala jooksul. Härra Hani luges ühe nädala jooksul kokku kümne kana munad: 3 5 4 6 2 6 5 6 5 3.
95%-line usaldusintervall = ?

$$\bar{x} = 4,5;\ s \approx 1,43 \qquad t_{1-\alpha/2;(n-1)} = t_{0,975;9} = 2,26$$

$$\underline{\mu};\bar{\mu} = \left(\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = \left(4,5 - 2,26\frac{1,43}{\sqrt{10}}; 4,5 + 2,26\frac{1,43}{\sqrt{10}}\right)$$

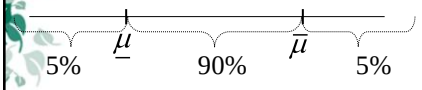
$$= 4,5 - 2,26 \times 0,45;\ 4,5 + 2,26 \times 0,45 = 3,47;\ 5,53$$

95%-lise tõenäosusega võib väita, et keskmine nädalas munetud munade arv on kuskil vahemikus 3,47-st 5,53-ni.

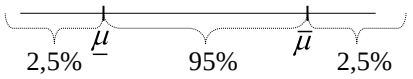
Suurendamaks hinnangu täpsust, tuleks uurida rohkem kanu, sest mida suurem on $n$, seda kitsamaks muutub usaldusintervall.

# Usaldusintervall

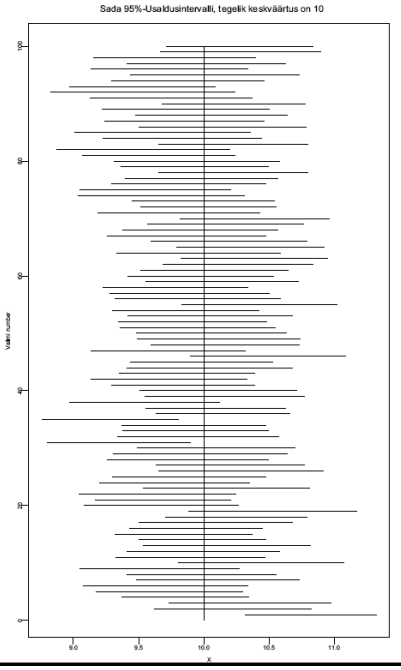
$\alpha$=0,1 → 90%-usaldusintervall




$\alpha$=0,05 → 95%-usaldusintervall




$\alpha$=0,01 → 99%-usaldusintervall


# Usaldusintervall



TABLE I. Values of rainbow trout blood chemistry values ($n$ = 45)

| | Mean ± S.E. | Median | Skewness (mean ± S.E.) | Kurtosis (mean ± S.D.) | CV (%) | 95% CI (lower–upper) |
|---|---|---|---|---|---|---|
| Glucose (mg dl$^{-1}$) | 108·11 ± 9·98 | 98·00 | 0·62 ± 0·35 | −0·21 ± 0·70 | 61·91 | 88·00–128·22 |
| Urea* (mg dl$^{-1}$) | 4·36 ± 0·24 | 4·00 | 1·21 ± 0·35 | 2·24 ± 0·70 | 36·35 | 3·88–4·83 |
| Creatinine* (mg dl$^{-1}$) | 0·29 ± 0·01 | 0·29 | 1·24 ± 0·35 | 3·09 ± 0·70 | 23·71 | 0·27–0·31 |
| Total bilirubin (mg dl$^{-1}$) | 0·04 ± 0·00 | 0·05 | 0·11 ± 0·35 | −0·55 ± 0·70 | 58·02 | 0·04–0·05 |
| Aspartate aminotransferase (U l$^{-1}$) | 461·20 ± 27·62 | 447·00 | 0·65 ± 0·35 | 0·38 ± 0·70 | 40·18 | 405·53–516·87 |
| Alanine aminotransferase* (U l$^{-1}$) | 12·87 ± 1·16 | 11·00 | 1·71 ± 0·35 | 3·63 ± 0·70 | 60·22 | 10·54–15·19 |
| Alkaline phosphatase* (U l$^{-1}$) | 179·22 ± 19·26 | 131·00 | 1·92 ± 0·35 | 3·84 ± 0·70 | 72·10 | 140·40–218·04 |
| Creatine phosphokinase* (U l$^{-1}$) | 1265·11 ± 161·70 | 894·00 | 1·21 ± 0·35 | 0·73 ± 0·70 | 85·74 | 939·22–1591·00 |
| Lactate dehydrogenase* (U l$^{-1}$) | 2628·18 ± 164·75 | 2399·00 | 1·53 ± 0·35 | 2·62 ± 0·70 | 42·05 | 2296·15–2960·21 |
| Gamma-glutamyl transferase (U l$^{-1}$) | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. |
| Total protein (g dl$^{-1}$) | 3·59 ± 0·13 | 3·74 | −0·59 ± 0·35 | 0·46 ± 0·70 | 24·64 | 3·32–3·85 |
| Albumin (g dl$^{-1}$) | 1·38 ± 0·05 | 1·40 | −0·39 ± 0·35 | 0·05 ± 0·70 | 25·17 | 1·27–1·48 |
| Triglycerides (mg dl$^{-1}$) | 347·51 ± 23·56 | 327·00 | 0·46 ± 0·35 | −0·73 ± 0·70 | 45·47 | 300·04–394·99 |
| Cholesterol (mg dl$^{-1}$) | 247·38 ± 10·32 | 241·00 | 0·09 ± 0·35 | 0·10 ± 0·70 | 27·98 | 226·59–268·17 |
| Ca (mg dl$^{-1}$) | 12·52 ± 0·20 | 12·20 | 0·55 ± 0·35 | −0·08 ± 0·70 | 10·81 | 12·11–12·93 |
| P (mg dl$^{-1}$) | 22·66 ± 1·19 | 22·50 | 0·45 ± 0·35 | −0·67 ± 0·70 | 35·26 | 20·26–25·06 |
| Mg (mg dl$^{-1}$) | 3·85 ± 0·11 | 3·82 | 0·07 ± 0·35 | −0·84 ± 0·70 | 15·40 | 3·63–4·07 |
| Na (mEq l$^{-1}$) | 154·07 ± 0·85 | 155·00 | −0·53 ± 0·35 | 1·14 ± 0·70 | 3·69 | 152·36–155·78 |
| K (mEq l$^{-1}$) | 3·45 ± 0·29 | 3·25 | 0·45 ± 0·35 | −0·81 ± 0·70 | 52·55 | 2·87–4·03 |
| Cl* (mEq l$^{-1}$) | 128·09 ± 1·13 | 130·00 | −2·90 ± 0·35 | 11·84 ± 0·70 | 5·90 | 125·82–130·36 |

n.a., not assessable. *Null hypothesis (Kolmogorov–Smirnov test) was rejected.

# Hüpoteeside testimine

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

## Hüpoteeside kontroll

**Näiteid hüpoteesidest**

- ✓ Kas jogurti toiduvärviga värvimine parandab tarbijate meelest selle maitseomadusi?
- ✓ Kas leidub seos lehma tiinestumise ja piimatoodangu vahel?
- ✓ Kas nn õnnelike sigade tailiha % on erinev tavalises sigalas kasvanud sigade vastavast näitajast?
- ✓ Kas Eesti ja Soome vetest püütud lõhed on geneetiliselt erinevad?

**Hüpoteeside paar**

$H_1$ – väide, mida me soovime tõestada (sisukas e alternatiivne hüpotees; *alternative hypothesis*),

$H_0$ – väide, et üldkogum vastab teatavale standardile (nullhüpotees; *null hyphothesis*).

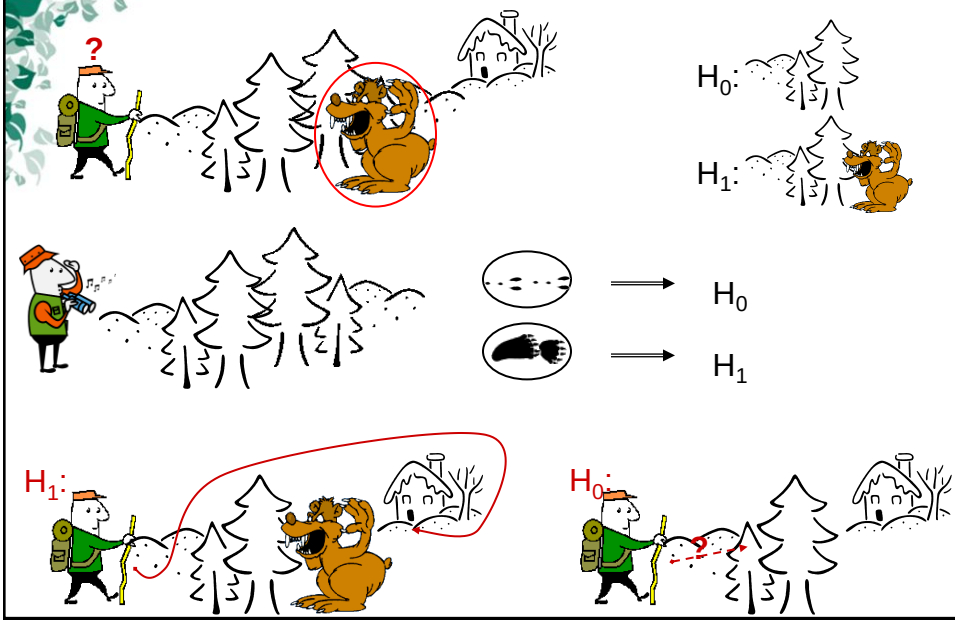
**Teststatistik** – valimifunktsioon, mis mõõdab erinevust nullhüpoteesis väidetu ja andmetest ilmneva vahel.

## Hüpoteeside kontrolli loogika

**Hüpoteeside kontrolli algoritm**

1. Eeldame, et populatsioonis kehtib nullhüpotees.
2. Arvutame andmete (valimi) alusel teststatistiku väärtuse.
3. Kui saadud teststatistiku väärtus on nullhüpoteesi eeldusel väga ebatõenäoline, on loomulik järeldada, et algne eeldus nullhüpoteesi kehtimise kohta oli vale, ning lugeda tõestatuks alternatiivne hüpotees.
4. Kui teststatistiku väärtus jääb piiridesse, kus ta nullhüpoteesi kehtides ka peaks olema, ei ole mingit alust nullhüpoteesi ümber lükata (samas ei saa nullhüpoteesi ka tõestatuks lugeda, sest äkki vastavad uuringu tulemused nullhüpoteesile lihtsalt juhuslikult!).

## Hüpoteeside kontrolli loogika



## Hüpoteeside kontroll

**Vead hüpoteeside kontrollimisel**

Esimest liiki viga tekib siis, kui võetakse vastu sisukas hüpotees, aga tegelikult on õige nullhüpotees.

Teist liiki viga tekib siis, kui jäädakse nullhüpoteesi juurde, kuid õige oleks sisukas hüpotees.

| Tegelik olek / Otsus | Õige $H_0$ | Õige $H_1$ |
|---|---|---|
| Jääme $H_0$ juurde | + | II liiki viga, $\beta$ |
| Kummutame $H_0$ | I liiki viga, $\alpha$ | + |

**Olulisuse nivoo** $\alpha$ (*significance level*) – maksimaalne lubatav I liiki vea tõenäosus (tavaliselt $\alpha$ = 0,05; 0,01; 0,001), nö valulävi.

**Testi võimsus** [*power*] = 1–$\beta$ on tõenäosus lugeda õigeks ka tegelikult kehtiv sisukas hüpotees $H_1$.

## Hüpoteeside kontroll

**Olulisuse tõenäosus** (***p*-väärtus**; *probability level, p-value*)

– tõenäosus eksida, väites oma andmete põhjal sisuka hüpoteesi $H_1$ kehtimist (I liiki vea tegemise tõenäosus);

– tõenäosus saada analüüsitava struktuuriga ("nii suure erinevusega" või "nii tugeva seosega") andmed juhuslikult – P(valim|$H_0$).

**Otsuse vastuvõtmine (1)**

Võrreldakse olulisuse tõenäosust $p$ ja olulisuse nivood $\alpha$:

¤ **kui $p \le \alpha$, siis on tõestatud $H_1$,**

¤ **kui $p > \alpha$, siis jääme $H_0$ juurde**.

*p*-väärtus leitakse kas tuginevalt teoreetilistele jaotustele või permutatsioonimeetodil.

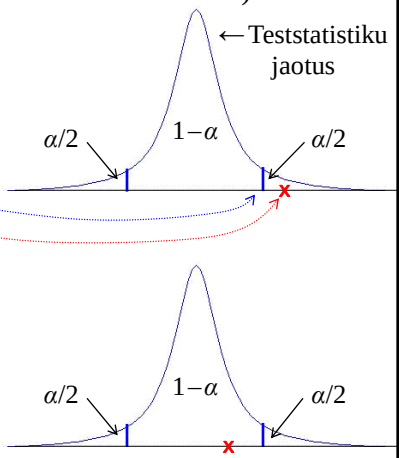
## Hüpoteeside kontroll

**Otsuse vastuvõtmine (2)**

Võrreldakse arvutatud teststatistiku väärtust selle kriitilise väärtusega (tuginedes teoreetilistele jaotustele või simuleerimise tulemustele):

¤ kui teststatistiku absoluutväärtus on suurem tema nullhüpoteesipõhise jaotuse kriitilisest väärtusest (1–$\alpha$/2-kvantiilist), loetakse õigeks $H_1$,

¤ vastupidisel juhul jäädakse nullhüpoteesi $H_0$ juurde.



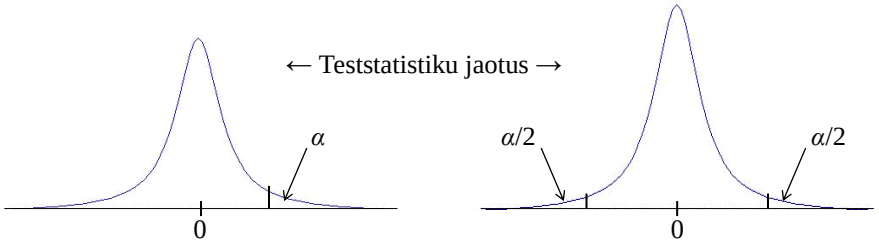

## Hüpoteeside kontroll

**Ühepoolne** [*one-tail*] ***versus*** **kahepoolne** [*two-tail*] **hüpotees**

Näiteks:

$H_0 : \mu_1 > \mu_2$
$H_1 : \mu_1 \leq \mu_2$

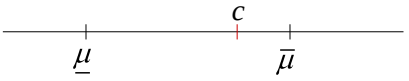
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$



## Seos hüpoteeside kontrolli ja usalduspiiride vahel

$H_0 : \mu = c$

võetakse vastu siis, kui $c$ kuulub usalduspiirkonda




$H_1 : \mu \neq c$

on tõestatud siis, kui $c$ ei kuulu usalduspiirkonda (olulisusnivool $\alpha$)


Praktikas kontrollitakse sageli kas mingi kordaja või gruppide vahe erinevust nullist.
Erinevuse võib lugeda statistiliselt oluliseks ette antud olulisuse nivool (näiteks $\alpha$=0,05), kui uuritavale kordajale või võrreldavate gruppide erinevusele (näiteks $\mu_1-\mu_2$) konstrueeritud (95%-line) usaldusintervall ei sisalda nulli!

## Keskväärtuse võrdlemine konstandiga

**Näide.** Kümme sassexi tõugu kana munesid nädalas vastavalt 3, 5, 4, 6, 2, 6, 5, 6, 5 ja 3 muna. Teades, et njuuhämpširi tõugu kanad munevad keskmiselt 5,4 muna nädalas, kontrollida hüpoteesi kahe tõu munatoodangute erinevusest.

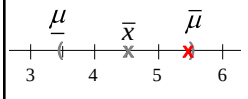
$H_0 : \mu = 5{,}4$
$H_1 : \mu \neq 5{,}4$

$n = 10;\ \bar{x} = 4{,}5;\ s \approx 1{,}43$

95%-lised usalduspiirid keskmisele nädalasele munatoodangule:

$$\underline{\mu}; \bar{\mu} = \left( \bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \right) = \left( 4{,}5 - 2{,}26\frac{1{,}43}{\sqrt{10}}; 4{,}5 + 2{,}26\frac{1{,}43}{\sqrt{10}} \right)$$

$$= 4{,}5 - 2{,}26 \times 0{,}45; 4{,}5 + 2{,}26 \times 0{,}45 = 3{,}47; 5{,}53$$




Järeldused: et $\underline{\mu} = 3{,}47 < 5{,}4 < 5{,}53 = \bar{\mu}$,
siis ei ole meil olulisuse nivoo $\alpha = 0{,}05$ korral alust ümber lükata nullhüpoteesi sellest, et sassexi tõugu kanad munevad sama palju kui njuuhämpširi tõugu kanad.

# Keskväärtuse võrdlemine konstandiga

**Usalduspiirid**

$H_0 : \mu = c$


võetakse vastu siis, kui $c$ kuulub usalduspiirkonda

$H_1 : \mu \neq c$


on tõestatud siis, kui $c$ ei kuulu usalduspiirkonda (olulisusnivool $\alpha$)

**Normaaljaotuse eeldusel *t*-test**

Teststatistik: $t = \frac{\bar{x} - c}{s}\sqrt{n} \underset{H_0}{\sim} t_{n-1}$

$|t| \geq t_{1-\alpha/2,n-1} \quad <=> \quad p \leq \alpha \quad \Rightarrow \quad H_1 : \mu \neq c$

$|t| < t_{1-\alpha/2,n-1} \quad <=> \quad p > \alpha \quad \Rightarrow \quad H_0 : \mu = c$

**Suurte valimite ($n$>60) korral *z*-test**

Teststatistik:

$Z = \frac{\bar{x} - c}{s}\sqrt{n} \underset{H_0}{\sim} N(0,1)$

$|Z| \geq z_{1-\alpha/2} \quad <=> \quad p \leq \alpha \quad \Rightarrow \quad H_1 : \mu \neq c$

$|Z| < z_{1-\alpha/2} \quad <=> \quad p > \alpha \quad \Rightarrow \quad H_0 : \mu = c$

# Keskväärtuse võrdlemine konstandiga

**Näide.** Kümme sassexi tõugu kana munesid nädalas vastavalt 3, 5, 4, 6, 2, 6, 5, 6, 5 ja 3 muna. Teades, et njuuhämpširi tõugu kanad munevad keskmiselt 5,4 muna nädalas, kontrollida hüpoteesi kahe tõu munatoodangute erinevusest.

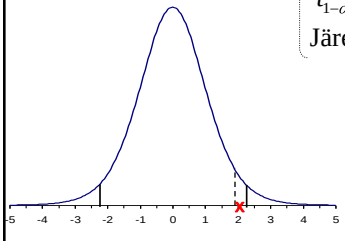
$H_0 : \mu = 5{,}4$
$H_1 : \mu \neq 5{,}4$
või
$H_0 : \mu \geq 5{,}4$
$H_1 : \mu < 5{,}4$

$n = 10;\ \bar{x} = 4{,}5;\ s \approx 1{,}43;\ \alpha = 0{,}05$

Teststatistik: $|t| = \left|\frac{\bar{x} - 5{,}4}{s}\sqrt{n}\right| = \left|\frac{4{,}5 - 5{,}4}{1{,}43}\sqrt{10}\right| = |-1{,}985| = 1{,}985$

Teststatistiku kriitiline väärtus (kahepoolne hüpotees):
$t_{1-\alpha/2;(n-1)} = t_{0{,}975;9} = 2{,}26$
Järeldus: $|t| = 1{,}985 < 2{,}26 = t_{0{,}975;9} \Rightarrow H_0 : \mu = 5{,}4$

Teststatistiku kriitiline väärtus (ühepoolne hüpotees):
$t_{1-\alpha;(n-1)} = t_{0{,}95;9} = 1{,}83$
Järeldus: $t_{0{,}95;9} = 1{,}83 < 1{,}985 = |t| \Rightarrow H_1 : \mu < 5{,}4$




Näiteks *MS Excelis*
funktsioon TDIST(t;n-1;2)

Arvuti abil saab leida ka täpse tõenäosuse teststatistiku väärtuse $|t| = 1{,}985$ saamiseks eeldusel, et kehtib $H_0$:
$p = 0{,}0784$ (2-poolne hüp.); $p = 0{,}0392$ (1-poolne hüp.)