Valim, üldkogum, hinnangud ja usaldusvahemikud

Krista Fischer

Üks põhilisi statistilise andmeanalüüsi ülesandeid on…

- Järelduste tegemine väga suure üldkogumi kohta kasutades selleks lõpliku suurusega valimit

Mõisteid

Üldkogum – kõigi huvipakkuvate objektide hulk (ka: populatsioon, rahvastik, rahvastikurühm, …)

Valim – uurimiseks (andmeanalüüsiks) kasutatav hulk objekte üldkogumist

Juhuvalim – valim, mis on saadud juhusliku valiku teel üldkogumist

Juhuvalimi vastand on käepärane valim, mis saadakse mittejuhusliku valiku teel, nii nagu uurijale mugav on

Miks on vaja juhuvalimit?

- Ainult juhuvalimi korral võib olla kindel, et valimi põhjal saadud hinnangud on nihketa hinnanguteks üldkogumi vastavale näitajale

Eksperiment – kui palju kaaluvad kokku 200 kommi?

1. Paku – mida sa arvad, kui palju kaalub kotitäis komme?

2. Võta valim suurusega 10 kommi
   a) Kaalu kommid ära ja pane paberile kirja 10 kommi kogukaal
   b) Korruta summa 20-ga, hindamaks terve kausitäie kommide kaalu

Tulemused

Hinnang silma järgi	Valimi kaal	Kiss-kiss komme
1500	85	3
800	76	4
1000	73	2
2000	107	2
2000	66	3
2500	66	3
950	79	2
2500	11	4
1500	81	3
1500	80	2
1600	104	0
3000	114	2
1500	85	5
2125	76	4
2000	108	2
2000	87	3
750	85	2
1500	70	3
1200	70	4
1750	104	1
1500	64	4
1250	87	3
2000	77	3
1500	75	3

Kokkuvõte – vt teisest failist (R-kommid.pdf)

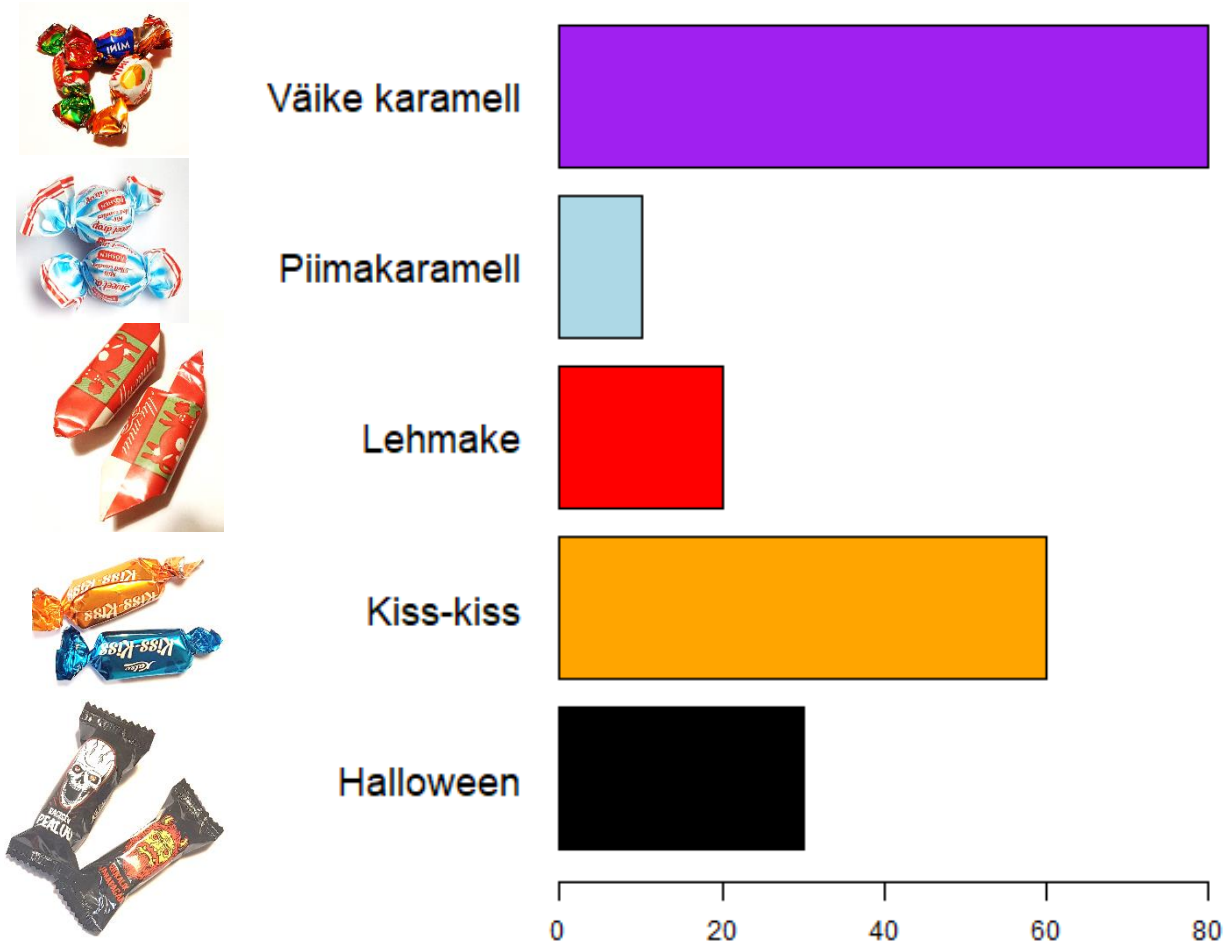
Milline on tegelik kommide jaotus?

Liik	Arv	%	Ühe kommi kaal (g)	Kogukaal
Halloween	30	15%	12,0	360
Lehmake	20	10%	18,0	361
Kiss-kiss	60	30%	4,3	257
Piimakaramell	10	5%	7,1	71
Väike karamell	80	40%	2,4	192
	200			**1241**

Keskmine 10 kommiga juhuvalimi kaal võiks olla 1241/20 =**62,05g**

Mitu valimit ülehindas kommide kogukaalu (valimi kaal üle 62g)?
Mitu valimit alahindas (valimi kaal alla 62g)?

Kommide jaotus „üldkogumis“

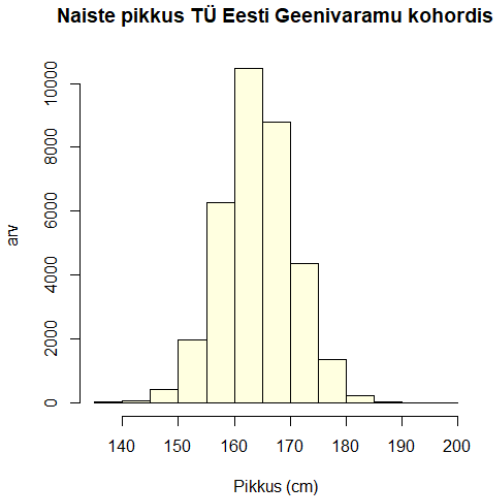


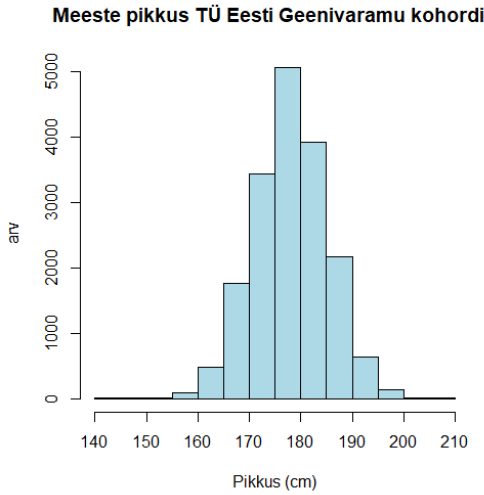
Tõenäosusjaotused *(probability distributions)* – mis nad on ja milleks neid on vaja?

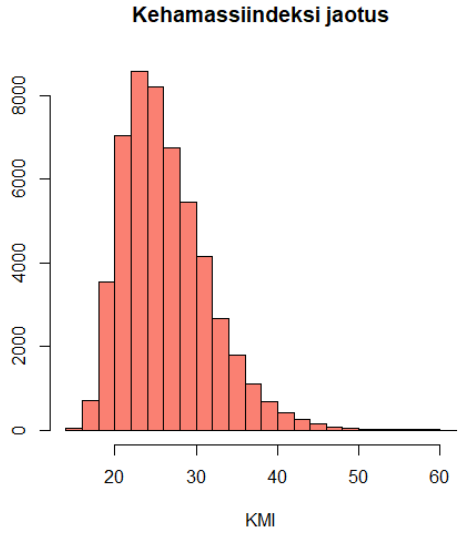
- Kui tunnuse jaotus (vaatluste jagunemine erinevate väärtuste vahel) on kirjeldatav teatud matemaatiliste valemite abil, saame öelda, et andmed vastavad teatud (parameetrilisele) jaotusele

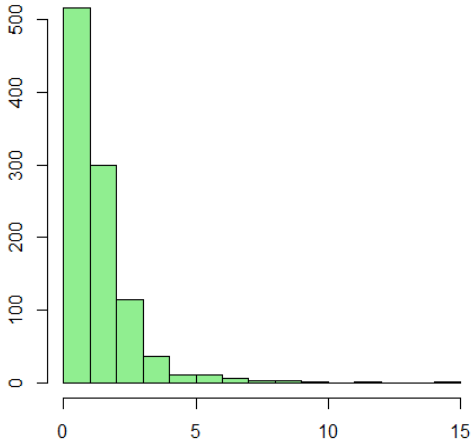
- Tüüpised pidevate tunnuste jaotused: normaaljaotus (Gaussi jaotus), lognormaaljaotus, exponentsiaaljaotus,..
- Diskreetsed andmed: binoomjaotus, Poissoni jaotus,…
- Mittearvulised tunnused: multinominaaljaotus, hüpergeomeetriline jaotus

Histogrammid andmetele

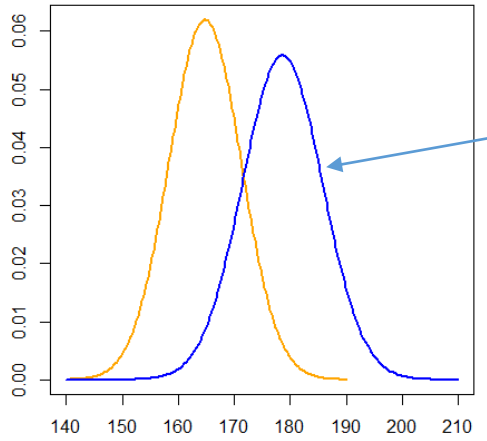






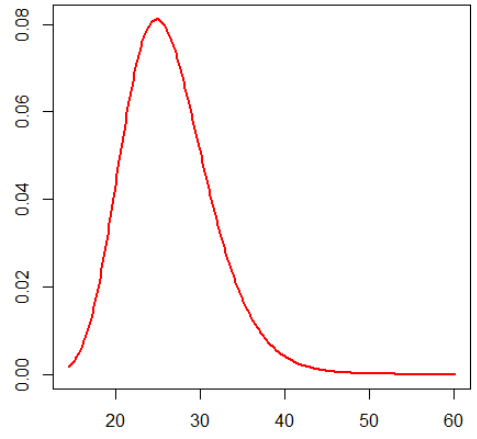


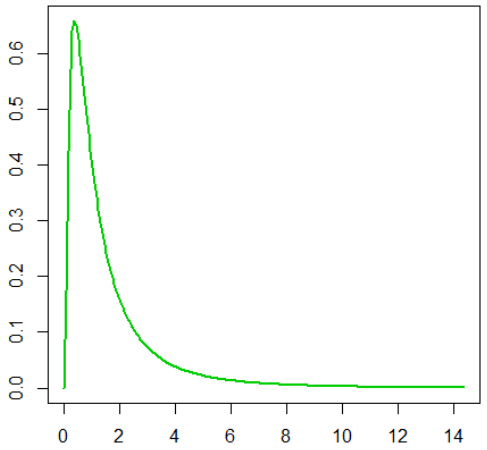
Teoreetilised jaotused



Matemaatiline funktsioon:

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





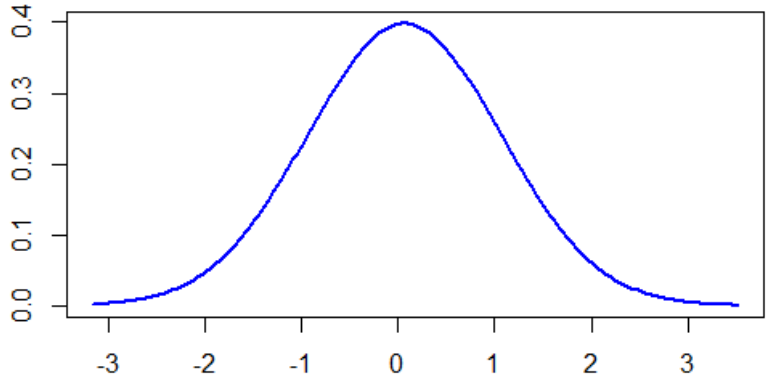
Miks me vajame jaotuseid?

- Enamus statistilisi meetodeid eeldab, et andmed järgivad teatud jaotust (vähemalt ligikaudu)

- Erijuhtudel, kus ükski parameetriline jaotus andmetega ei sobi, kasutatakse nn mitteparameetrilisi meetodeid

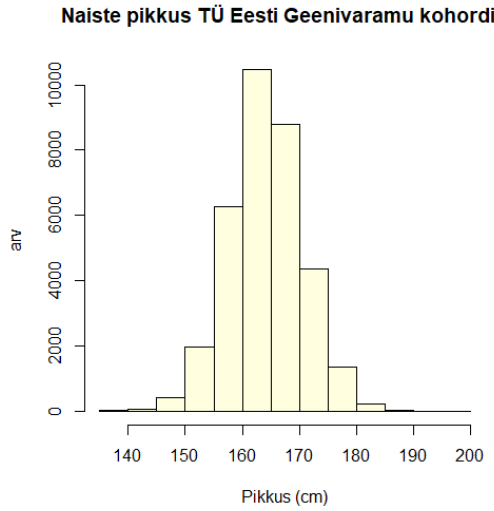
Normaaljaotus (*Normal distribution, Gaussian distribution*)

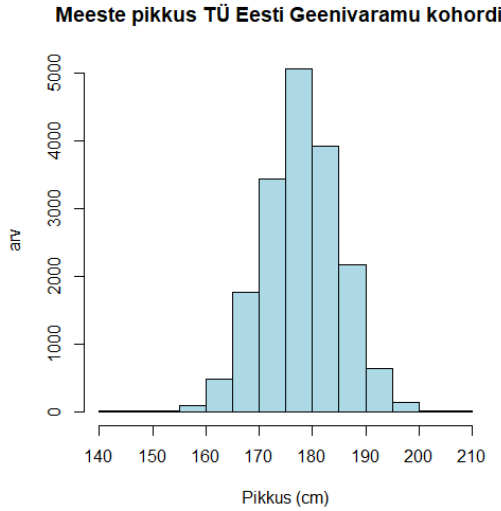
Pidev jaotus, mida iseloomustab sümmeetriline, nn kellukesekujuline tiheduskõver (histogramm) ehk Gaussi kõver.



Carl Friedrich Gauss
1777-1855

Normaaljaotusega andmed



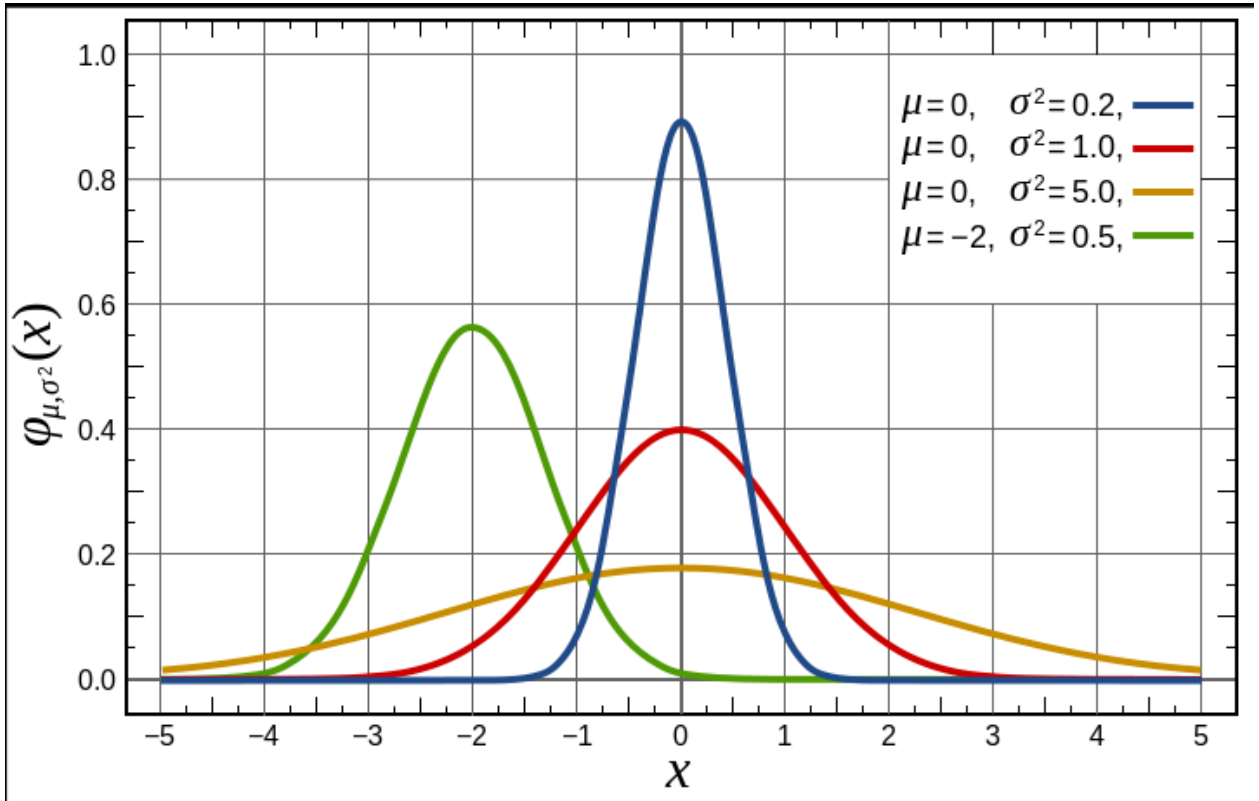


$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Kaks parameetrit: μ (keskväärtus) ja σ (standardhälve)

Kui andmed on normaaljaotusega, siis…

- Keskmine ja mediaan on ligikaudu võrdsed
- Ei esine suuri erindeid
- Andmeid võib analüüsida kõige tavalisemate statistikameetodite abil



Valimikeskmiste jt statistikute jaotus

- Võttes samast üldkogumist korduvalt valimeid, saame iga kord veidi erineva keskmise, protsendi vm näitaja
- Siiski, kõik valimikeskmised, protsendid jpt teised näitajad on ligikaudu normaaljaotusega, mille keskväärtuseks on üldkogumi keskväärtus

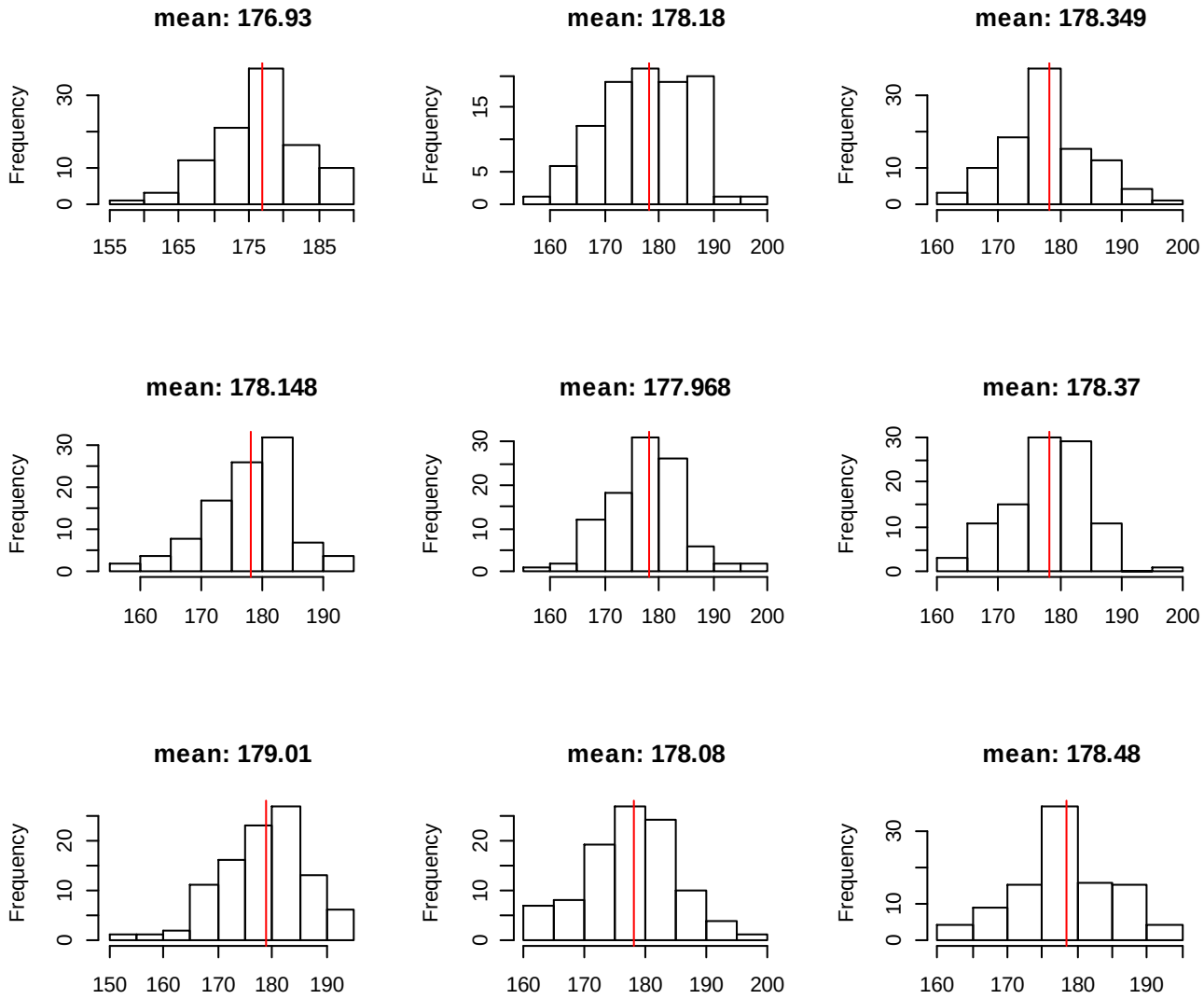
Valimikeskmise jaotus
(Sampling distribution of the mean)

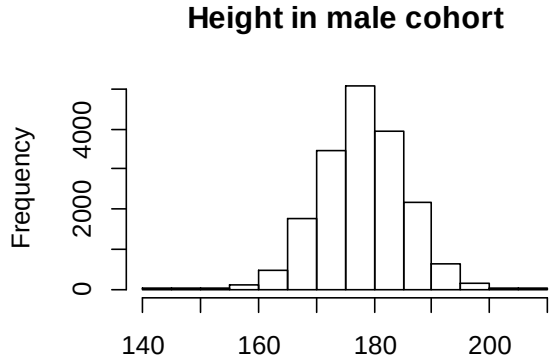
- Kui täpselt sama uuringut korrata erinevate valimitega samast üldkogumist, saame iga kord veidi erineva keskmise.
- Siiski, kõik saadud valimikeskmised on ligikaudu normaaljaotusega, mille keskväärtuseks on üldkogumi keskmine μ, ja standardhälbeks (ehk standardveaks, *standard error*) $\sigma/\sqrt{n}$, kus σ on üldkogumi standardhälve ja n on valimisuurus

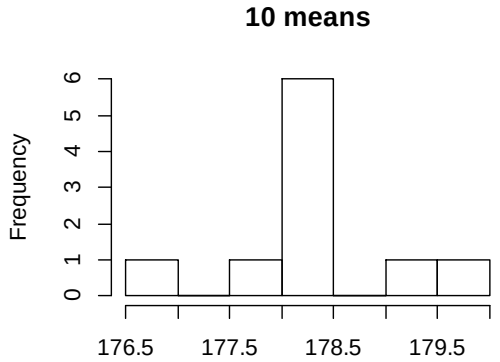
- Näide: meeste keskmine pikkus üldkogumis on 178.6cm, standardhälve 7.3. Seega võime oodata valimites suurusega 100, et valimikeskmine on keskmiselt 178.6, ja standardviga (valimikeskmiste standardhälve) 7.3/10=0.73

Veel üks näide

Valimid suurusega 100 (meeste pikkus)







Usaldusvahemiku (usaldusintervalli) mõiste

- 95% valimikeskmistest jäävad kuni 2 standardvea kaugusele üldkogumi keskmisest (vastavalt standardhälbe omadustele).

- Täpsemalt, kordaja 2 asemel tuleb tegelikult kasutada t-jaotuse kvantiili $t_{0.95;n-1}$, mis sõltub valimisuurusest, kuid on enamasti väga lähedal 2-le.

Vahemikku

$$\bar{x} \pm t_{0.95,n-1} \frac{s}{\sqrt{n}}$$

nimetatakse 95% usaldusvahemikuks (*confidence interval*) valimikeskmisele (95% CI).