Sissejuhatus statistika erialasse

loeng, 25. sept

Andmed

Andmekogumine, andmetüübid, andmete kirjeldamise ja visualiseerimise alused I

Krista Fischer

TÜ matemaatilise statistika professor

Tartu, 2019

Andmeid saab koguda mitmel erineval moel…

Typical example of an in-depth customer/patient/visitor survey

To be conducted by field surveyors in the form of an interview at a specific institution, hospital, school etc.

Ref Number (e.g. GIO 1234 P)	
Date (e.g. 01/12/2014)	
Surveyor name	

Surveyor Guidelines:

1. Please use clear and unadorned handwriting to assist with the ease of data capturing
2. Approach the respondent in a friendly, emotionally neutral and unbiased manner
3. Use the attached reference sheet of appropriate summaries for certain generic responses in the questions' comments section.
4. Leave a response blank if the respondent does not answer the question or responds with "I do not know."
5. If you have any questions, please contact your Project Manager, Peter, at 076 765 4567

Survey Part #1 - Respondent's Details

Gender	Male		Female		Unknown	
Age	13-18	19-29	30-39	40-60	60+	
Race						
Residential Area						

Survey Part #2 – Qualifying the Respondent

When was the last time you visited this institution?

How often do you visit this institution?	Last year	Last month	Last week	Very Recently
What is the state of your health at the moment? (If the institution is a Hospital)	Rarely or never	Not often	Now and then	Very often
	Very bad	Bad	Good	Very good

Survey Part #3 – The Institution

Respondent Guidelines:

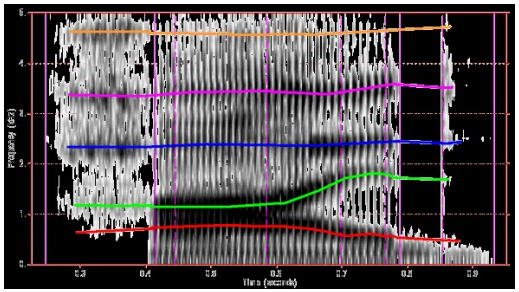
Please rate your level of agreement to every question on a scale of 1 to 4.

- 1 being Disagree;
- 2 being Somewhat Disagree;
- 3 being Somewhat Agree; and
- 4 being Totally Agree

Please indicate the response with an X in the corresponding field below.

No.	Question	1 - rating	2 - rating	3 - rating	4 - rating
#3.1	This institution is easy to access				
#3.2	Entrance and routes are clearly marked				
#3.3	The receptionist is friendly				
#3.4	Hours that the institution is open is acceptable				
#3.5	The waiting times and queues are acceptable				
#3.6	The level of comfort is acceptable				
#3.7	The level of safety is acceptable				
#3.8	Suggestion boxes are easily accessible				

Comments:



Oluline on, et andmekogumine oleks süstemaatiline ja hästi planeeritud

Küsimused, millele peaks vastama:

- Mis/kes on uuringuobjektid (inimesed, hiired, taimed, bakterid, tehingud, tooted…)?
- Kuidas koostatakse uuringuvalim? (juhuvalim, käepärane valim, …)
- Milliseid tunnuseid mõõdetakse/määratakse igal objektil?
- Mitu objekti kaasatakse uuringusse?
- Kas objektid on sõltumatud?
- Kas samadel objektidel teostatakse korduvaid mõõtmisi?

Uuringu planeerimise teemade juurde pöördume tagasi hiljem…

Andmestiku/andmefaili formaat

Sõltumata andmekogumise vormist ja sellest, milliseid andmeid kogutakse, on klassikalise andmeanalüüsi läbiviimiseks vaja organiseerida andmed andmetabeli ehk andmemaatriksi kujul

Andmemaatriks, kus on toodud 4 indiviidi sugu, vanus, kaal (kg) ja pikkus (cm) :

No	SUGU	VANUS	KAAL	PIKKUS	KOHV
1	Naine	57	65	157	0
2	Mees	70	100	176	3
3	Mees	45	91	181	1
4	Naine	38	58	160	5

Iga rida vastab ühele objektile ehk vaatlusele – siin andmestikus on 4 vaatlust

Iga veerg vastab ühele tunnusele – siin on kokku 5 tunnust

Kõik read on ühepikkused: kõigi indiviidide kohta on esitatud samad andmed.

Nõuded andmefailile (kehtivad enamike andmeanalüüsi tarkvarade kasutamisel)

- Iga rida vastab ühele vaatlusele – enamasti ühele objektile.
  Korduvmõõtmiste korral võib ühel objektil teostada ka mitu vaatlust.
- Iga veerg vastab ühele tunnusele.
  Mõnikord võib ühe mõõtmise või ankeedikküsimuse vastuse registreerimiseks olla vaja luua mitu tunnust (juhul kui tulemus ei ole kokku võetav ühe arvu või tekstistringi abil)
- Andmestiku ühes lahtris on ainult üks arv või tekstistring – väärtused, mida peab analüüsi käigus lugema võrdseks, peavad olema esitatud identselt
- Puuduvad väärtused on tihti paratamatud, kuid need tuleb andmestikus ära märkida üheselt arusaadaval moel – eri tarkvarapakettidel on siin erinevad nõudmised (NA, tühik, . , -99, jne)

Nõuded andmefailile: näide

Kood	Vanus liitumisel	Nohu või köha?
mehed		
A01234	34	Nohu
A09461	21,5	Nohu, köha
B98243	54	Ei ole
Naised		
C00001	23.5	Köha
C43278	45	Köha, nohu
C09375	21a, 6 kuud	Ei tea

Mis on siin valesti?

Tunnuste nimed ei pruugi olla tarkvara poolt loetavad

Vaheread mis ei vasta ühelegi vaatlusele, ei ole enamasti lubatud

Arvulise tunnuse puhul on kasutatud erinevat formaati – osa väärtuseid ei ole tarkvara poolt arvulisena loetavad

Sisuliselt identsed väärtused on kirja pandud erinevalt – neid loetakse erinevaks

Puuduva väärtuse tähistamine ei ole korrektne

Kuidas saaks paremini?

Kood	Vanus liitumisel	Nohu või köha?
mehed		
A01234	34	Nohu
A09461	21,5	Nohu, köha
B98243	54	Ei ole
Naised		
C00001	23.5	Köha
C43278	45	Köha, nohu
C09375	21a, 6 kuud	Ei tea

või

Kood	vanus	sugu	nohu	köha
A01234	34	1	1	0
A09461	21.5	1	1	1
B98243	54	1	0	0
C00001	23.5	2	0	1
C43278	45	2	1	1
C09375	21.5	2	NA	NA

Kuidas saaks paremini?

Kood	sugu	vanus	nohu	köha
A01234	1	34	1	0
A09461	1	21.5	1	1
B98243	1	54	0	0
C00001	2	23.5	0	1
C43278	2	45	1	1
C09375	2	21.5	NA	NA

- Lihtsad ja lühikesed (kuid arusaadavad) nimed tunnustele
- Sugu on andmestikus eraldi tunnusena – puudub vajadus andmestiku jagamiseks blokkidesse
- Sama formaat numbrite kirjapanekul
- Küsimuse „Nohu või köha?“ vastustest on hea moodustada kaks tunnust, mille väärtuste kirjapanekul saab kasutada lihtsat kodeeringut
- Puuduvad väärtused on tähistatud NA-ga

Veel andmesisestusest

- Andmesisestusel (nt küsitlusuuringu andmete korral) võib kasutada kodeeringut – sellisel juhul peab see aga olema hästi dokumenteeritud

Nt:

- Milline on teie haridus?
  1) Alg- või põhiharidus
  2) Keskharidus
  3) Kõrgharidus
  4) Teaduskraad

Andmestikus tunnus „haridus“, võimalike väärtustega 1, 2, 3 või 4, vastavalt vastaja tehtud valikule

Ülesanne

- Teile on antud lehekülg ühest küsimustikust – koostage andmemaatriks selle küsimustiku-osa andmete sisestamiseks (kas käsitsi paberile või nt Excelis)
  - Mitu tunnust on selles andmestikus?

- Sisestage (kirjutage) sinna andmestikku kahe isiku andmed (võivad olla väljamõeldud, kuid realistlikud)

- Milliseid probleemi/vigu võib tekkida nende andmete sisestamisel ja kuidas saaks neid vältida?

Kui Te olete kunagi tarbinud alkohoolseid jooke rohkem kui proovimiseks, siis kui vana Te olite, kui Te esimest korda jõite vähemalt 0,5 l õlut, 100 ml veini või 40 ml kanget alkoholi?

.....

Kui sageli Te keskmiselt alkohoolseid jooke tarvitate

- ☐ 4 või enamal korral nädalas
- ☐ 2–3 korda nädalas
- ☐ 1 korral nädalas
- ☐ 2–4 korda kuus
- ☐ korra kuus
- ☐ mõned korrad aastas
- ☐ vähem kui kord aastas

Kui sageli ja palju Te olete viimase 12 kuu jooksul tarvitanud järgmisi jooke

	Arv päevas	Või arv nädalas	Või arv kuus	Või arv aastas
Õlu, siider, long drink jms (0,5 l)				
Vein (100 ml)				
Kangestatud veinid, vermut, liköör (~20% vol) (40 ml)				
Kanged alkohoolsed joogid (40 ml)				

Kui Te olete kunagi tarbinud alkohoolseid jooke rohkem kui proovimiseks, siis kui vana Te olite, kui Te esimest korda jõite vähemalt 0,5 l õlut, 100 ml veini või 40 ml kanget alkoholi?

15

.....

Kui sageli Te keskmiselt alkohoolseid jooke tarvitate

- ☐ 4 või enamal korral nädalas
- ☒ 2–3 korda nädalas
- ☐ 1 korral nädalas
- ☐ 2–4 korda kuus
- ☐ korra kuus
- ☐ mõned korrad aastas
- ☐ vähem kui kord aastas

Kui sageli ja palju Te olete viimase 12 kuu jooksul tarvitanud järgmisi jooke

	Arv päevas	Või arv nädalas	Või arv kuus	Või arv aastas
Õlu, siider, long drink jms (0,5 l)		**4**		
Vein (100 ml)			**2**	
Kangestatud veinid, vermut, liköör (~20% vol) (40 ml)				
Kanged alkohoolsed joogid (40 ml)				**5**

Tunnuste tüübid

Arvulised tunnused

Üheselt mõistetava arvulise väärtusega (mis võib aga sõltuda kasutatavast mõõtühikust)

Pidevad tunnused

Pidev skaala: kaal, pikkus, temperatuur, …

167,3
-15

Diskreetsed tunnused

Väärtus määratakse loendamisel: laste arv, bakterite arv, vokaalide arv sõnas,…

2

Ka: kvantitatiivsed tunnused (*numeric data, quantitative data*)

Mittearvulised tunnused

Arvuline väärtus puudub (kuid andmestikus võib kasutada numbrilist kodeeringut)

Järjestustunnused

Järjestatud tasemed, hinnangud

Nominaaltunnused

Värv, rahvus, liik, elukoht, …

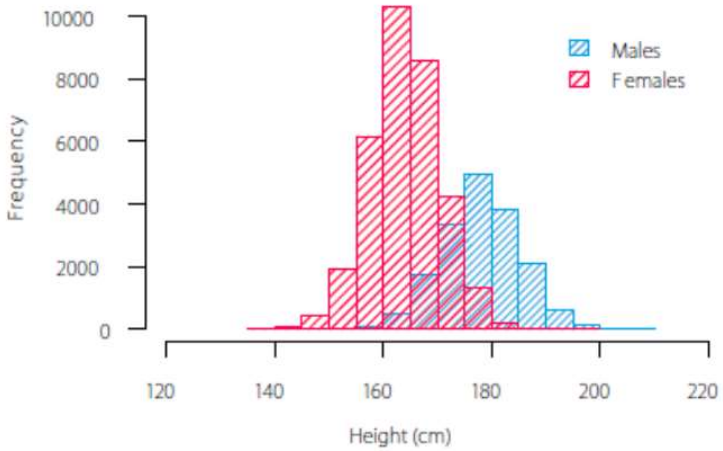
Ka: kvalitatiivsed tunnused, kategooriatunnused, grupitunnused (*categorical data*)

Erinevat tüüpi tunnused vajavad erinevaid meetodeid kirjeldamiseks ja visualiseerimiseks ning statistiliseks analüüsiks

- Pikkus

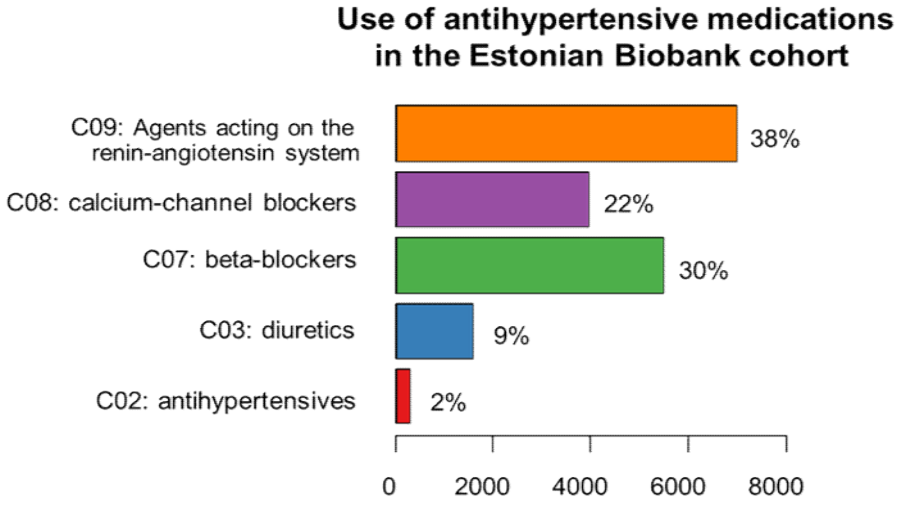
Keskmine eesti noormees oli aastal 1933 171,2 sentimeetrit pikk ning kaalus 67,38 kilogrammi. Selle teadasaamiseks mõõdeti ja kaaluti 7812 eesti sõdurit, ning see val

(Forte.delfi.ee, 02.08.2017)



(TÜ Eesti Geenivaramu)

- Vererõhuravimite tarbimine



Kirjeldav statistika I: mittearvulise tunnuse kirjeldamine

- Kõige tavalisem kokkuvõte andmetest: sagedustabel

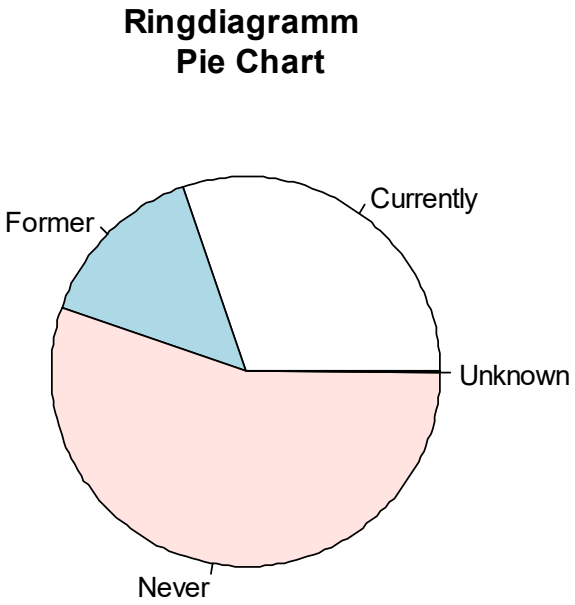
Suitsetamine	Praegune suitsetaja	Endine suitsetaja	Pole kunagi suitsetanud	Pole teada
N	11745	5617	21392	45
%	30,3%	14,5%	55,2%	0,1%

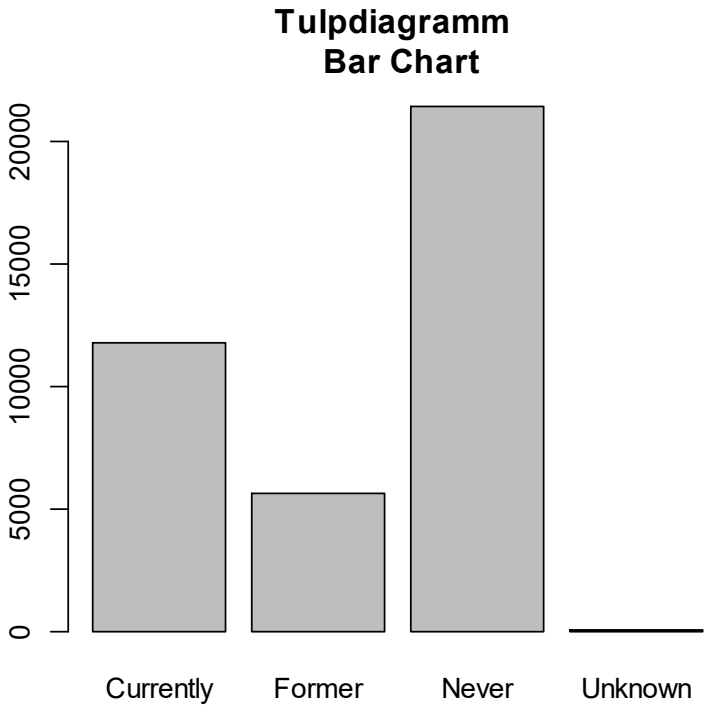
Mida tähele panna?

- Esitada alati lisaks protsendile ka absoluutsagedus (inimeste arv kategoorias)
- Protsendid tuleb enamasti ümardada – siin tuleks jälgida, et samas töös võiks üldiselt kasutada ühesugust täpsust arvandmete esitamisel
- Konkreetne formaat sõltub maitsest ja kirjutise vorminõuetest, kuid esmatähtis on informatiivsus

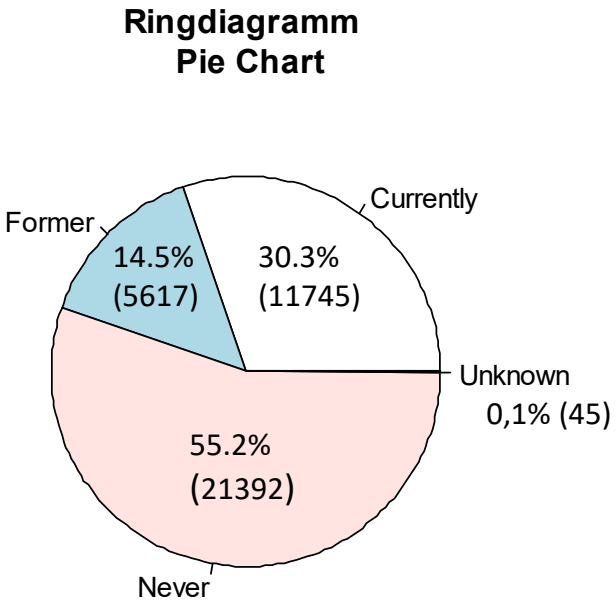
Kuidas seda graafiliselt kujutada?

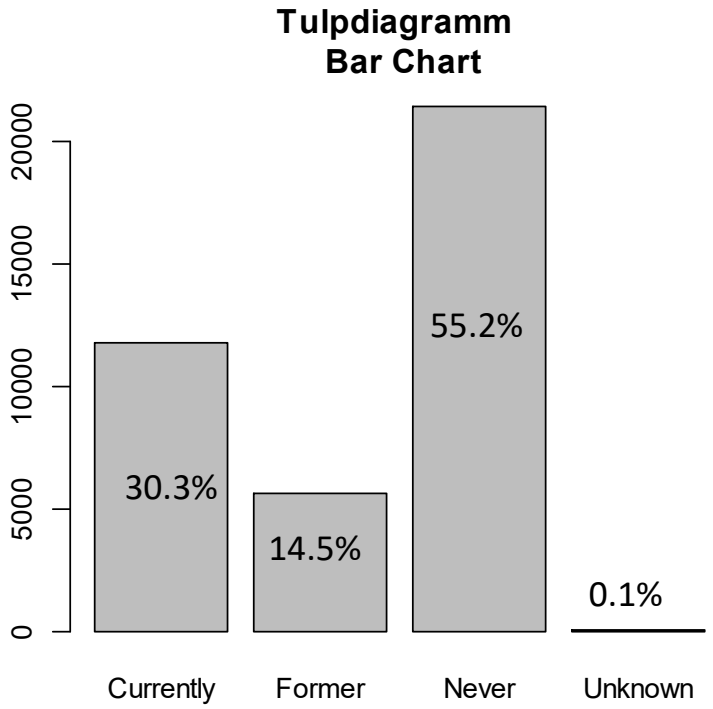
Kumb on parem esitus?





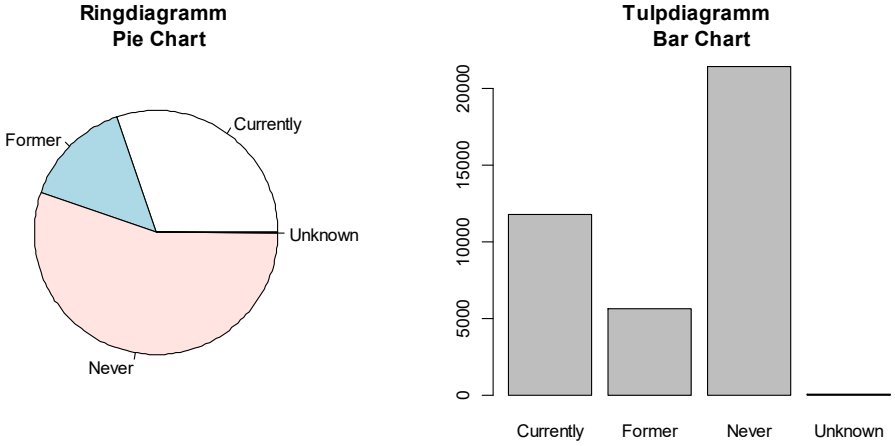
Kumb on parem esitus?

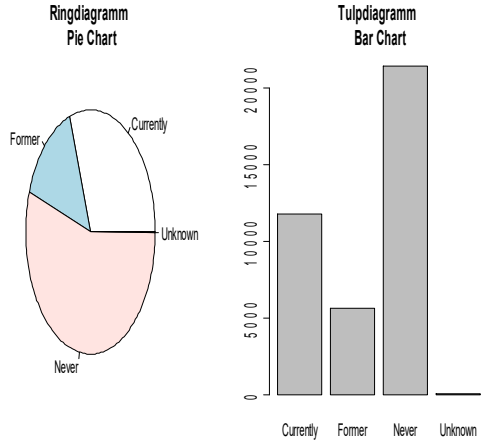




Kumb on parem esitus?

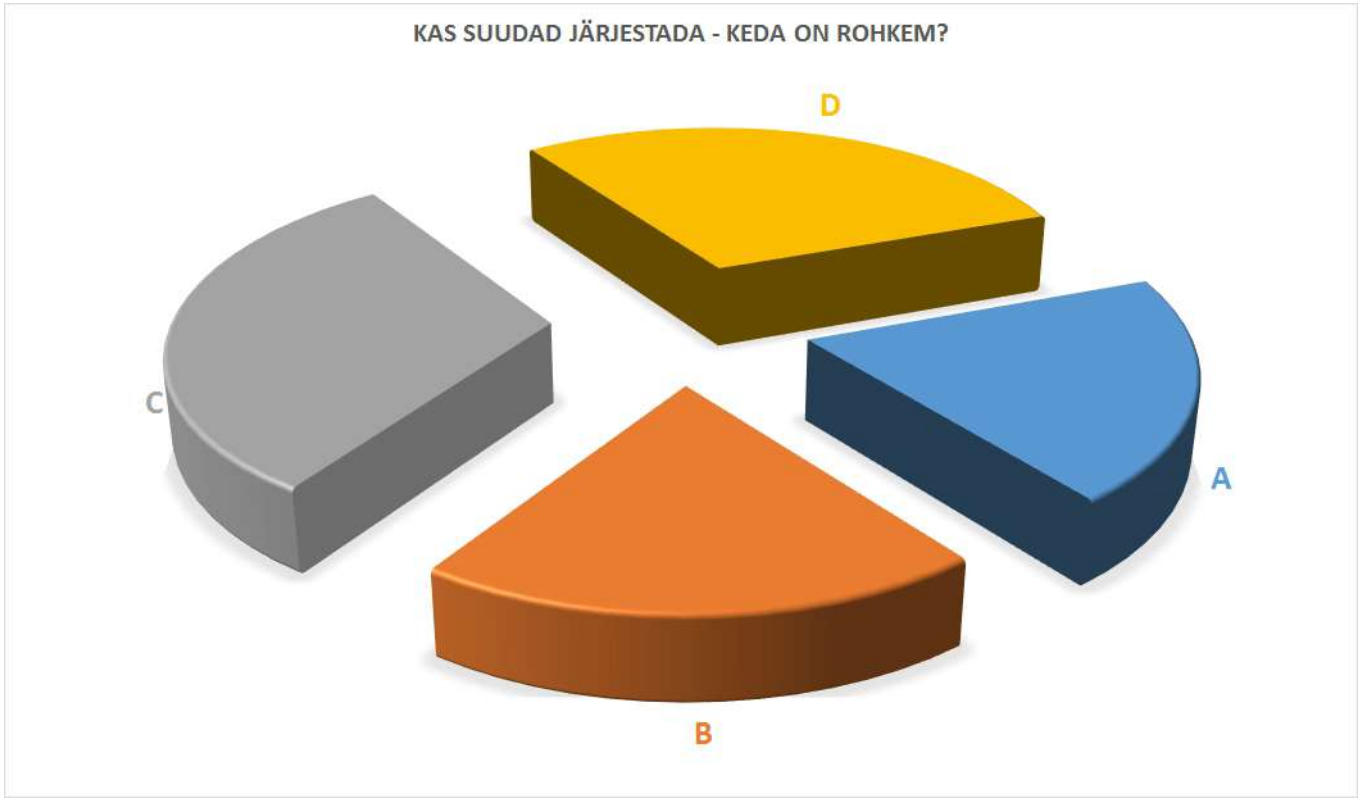
Enamikel juhtudel peetakse tulpdiagrammi sobivamaks – inimesel on lihtsam hinnata lineaarset skaalat kui võrrelda sektorite suuruseid

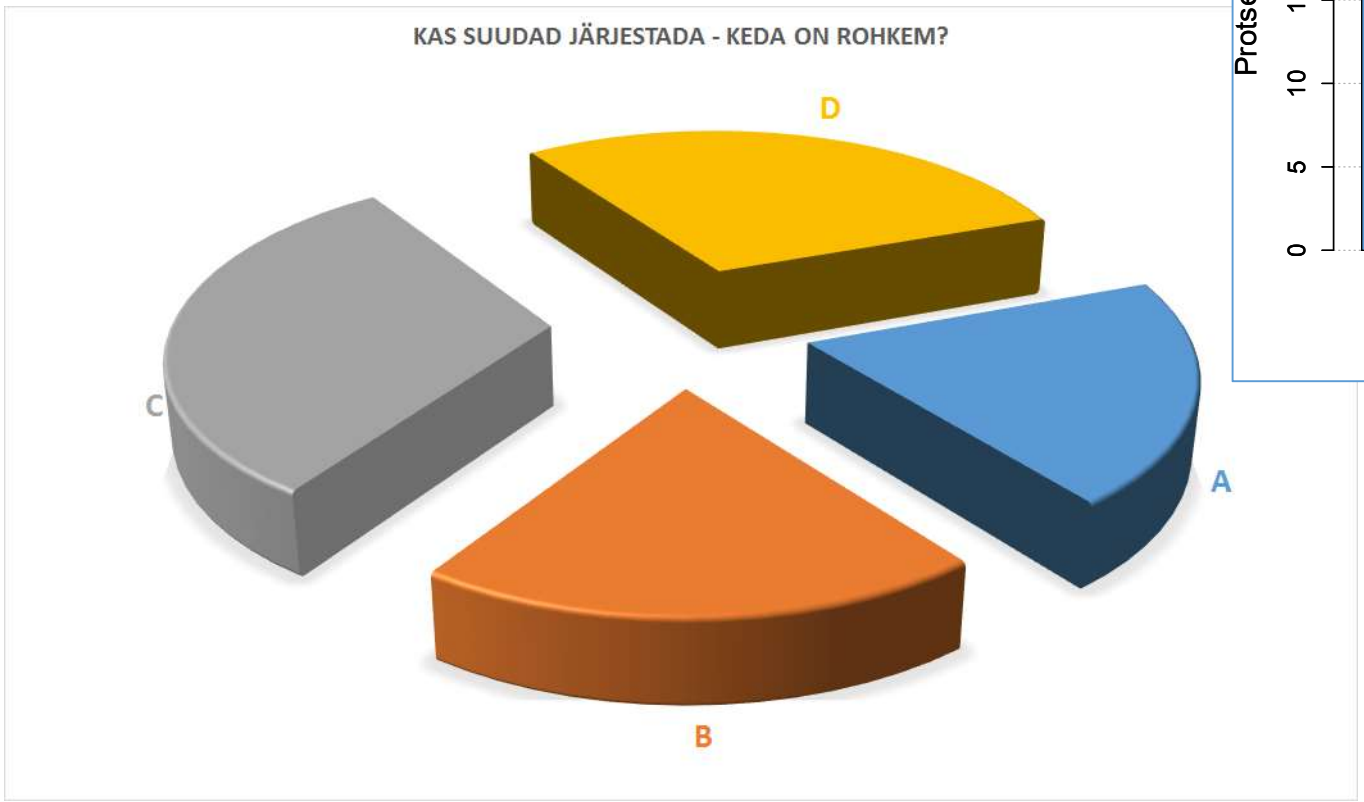


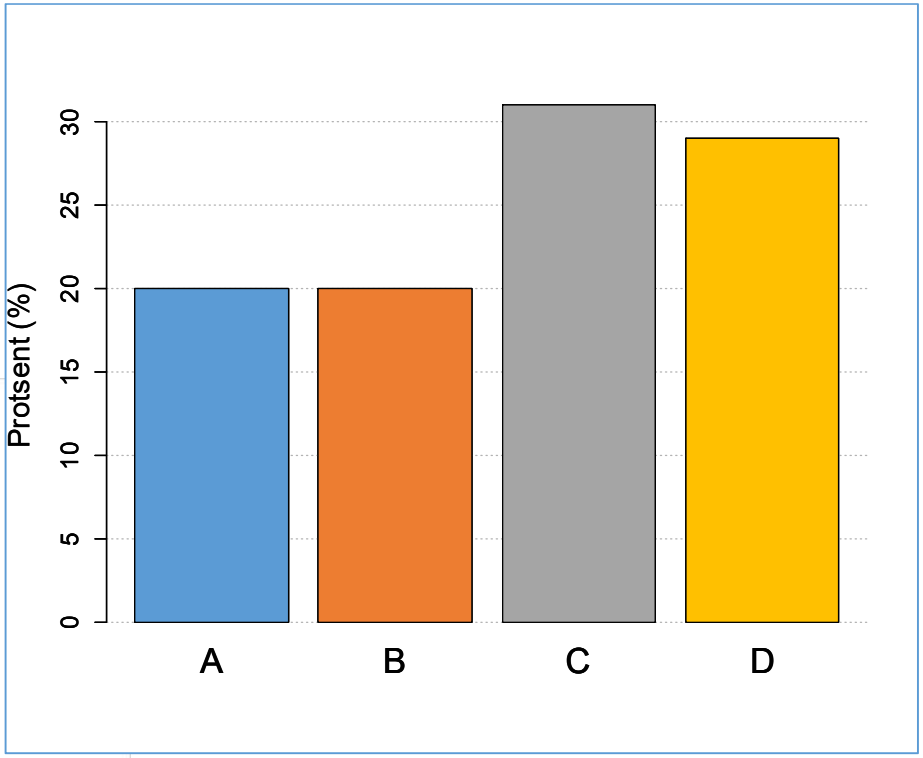


Eksitus / projektori ja arvuti mitte-ühilduvus….

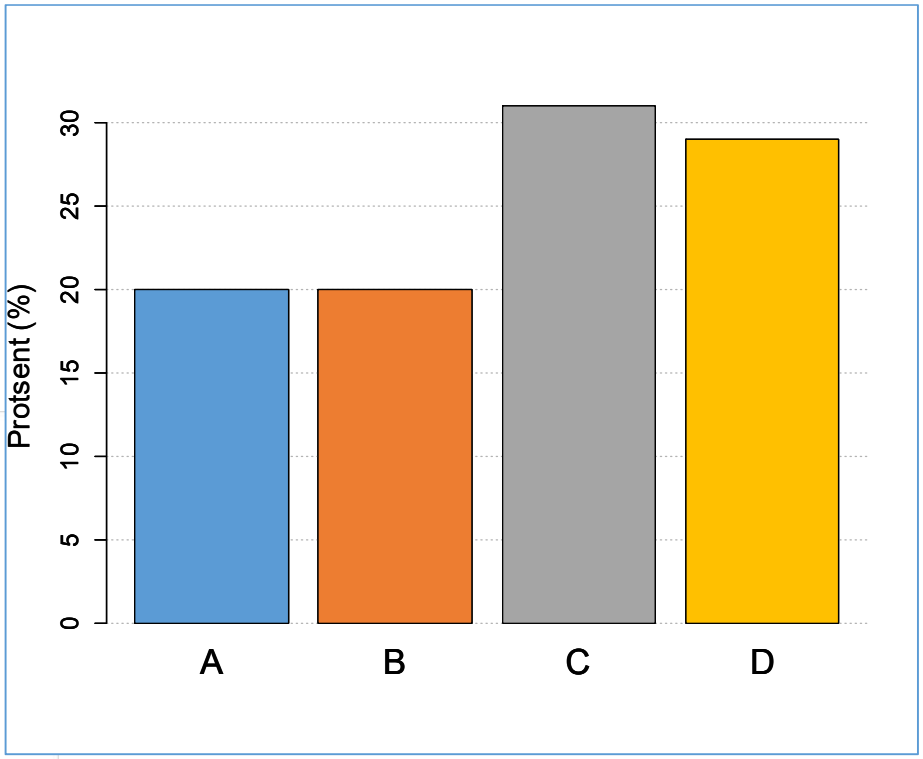
Ka kolmemõõtmelised pildid võivad tunduda uhked…



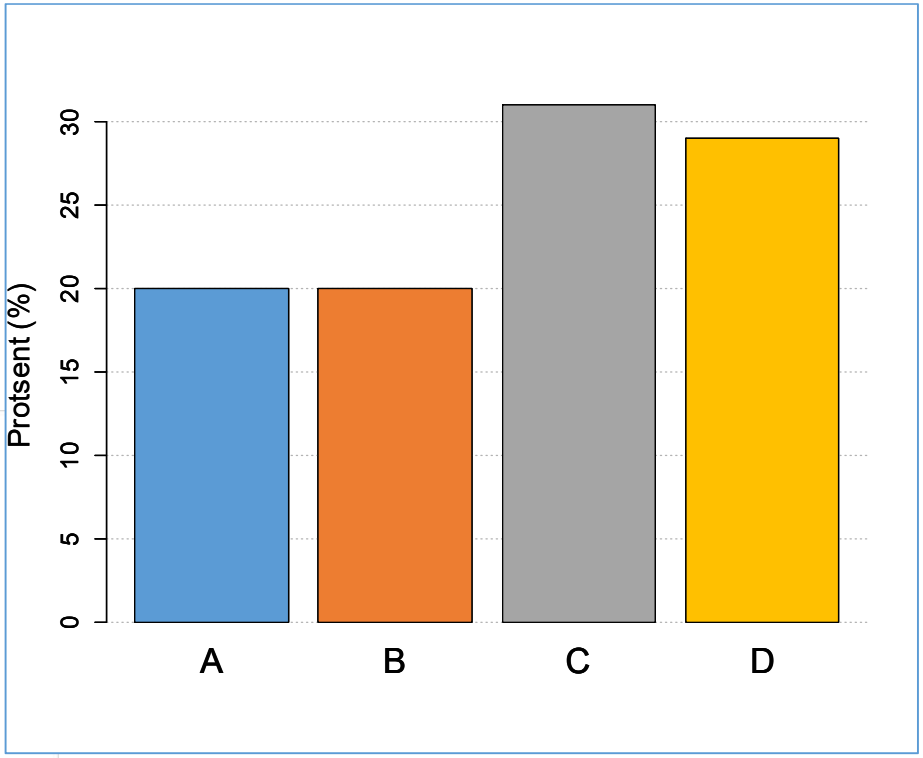




A	B	C	D
20	20	31	29

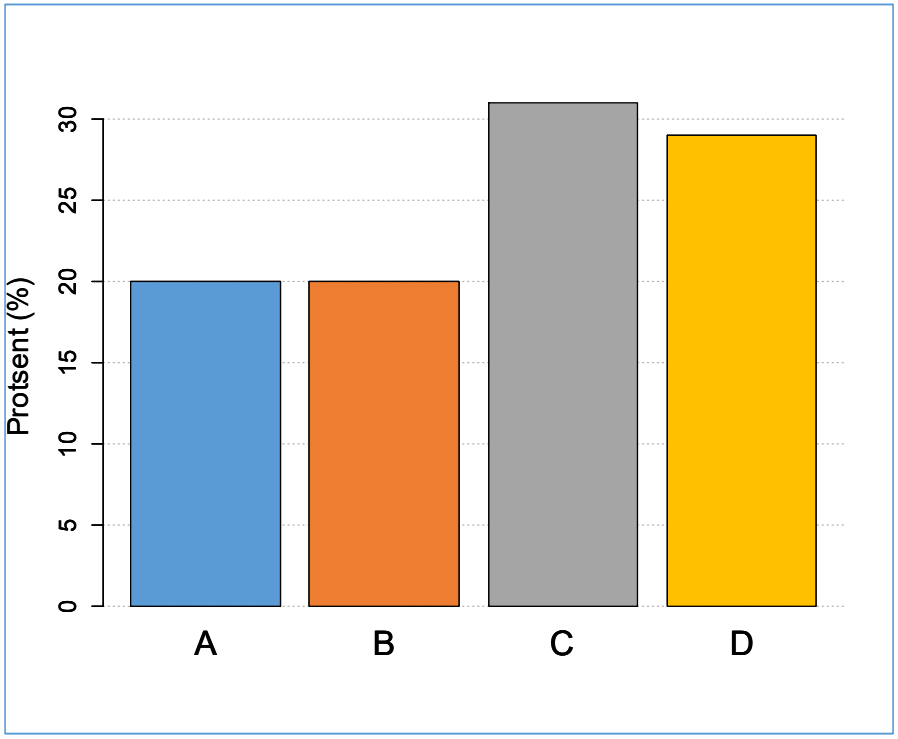


A	B	C	D
20	20	31	29



A	B	C	D
20	20	31	29

Lihtne ühetasandiline joonis (tulpdiagramm) võimaldab kõige paremini anda edasi kategooriate vahelisi erinevusi



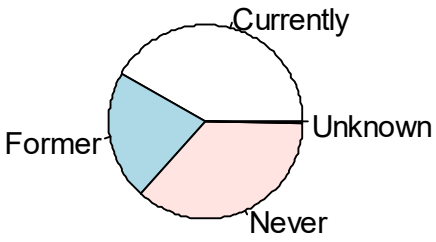
Kuidas lisada mõõtmeid?

Sugu / Suitsetamine	Mehed (n=12684)	Naised (n=26115)
Praegune	5312 (41,9%)	6433 (24,6%)
Endine	2717 (21,4%)	2900 (11,1%)
Mitte kunagi	4633 (36,5%)	16759 (64,2%)
Pole teada	22 (0,2%)	23 (0,1%)

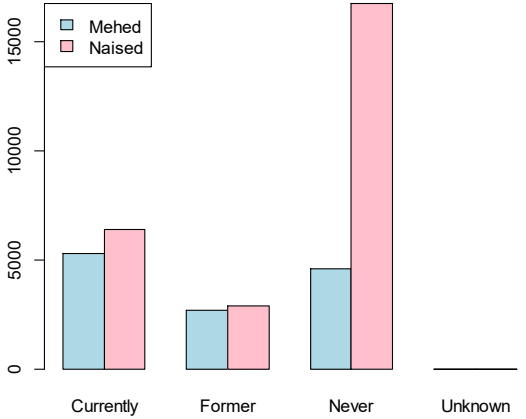
- Kas rea- või veeruprotsendid? Sõltub sellest, milliseid kategooriaid me võrrelda tahame

Kuidas nüüd graafiliselt kujutada?

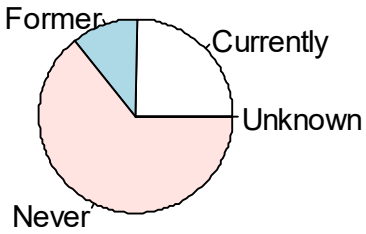
Suitsetamine: mehed

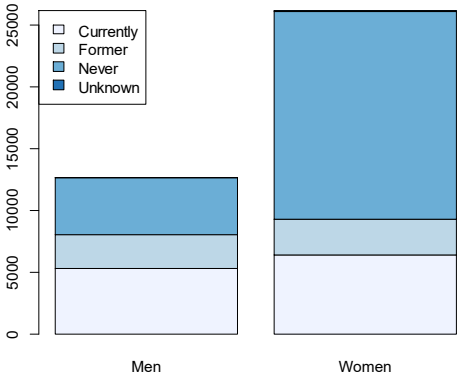


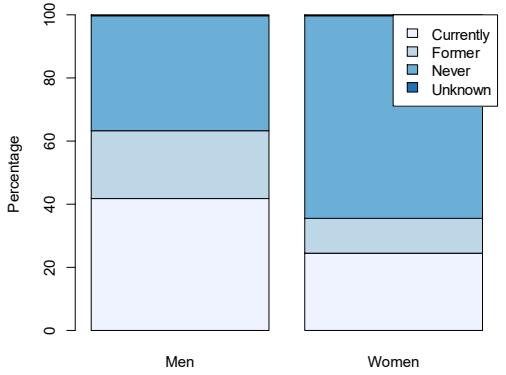
Lineaarne skaala on võrdluseks parem!



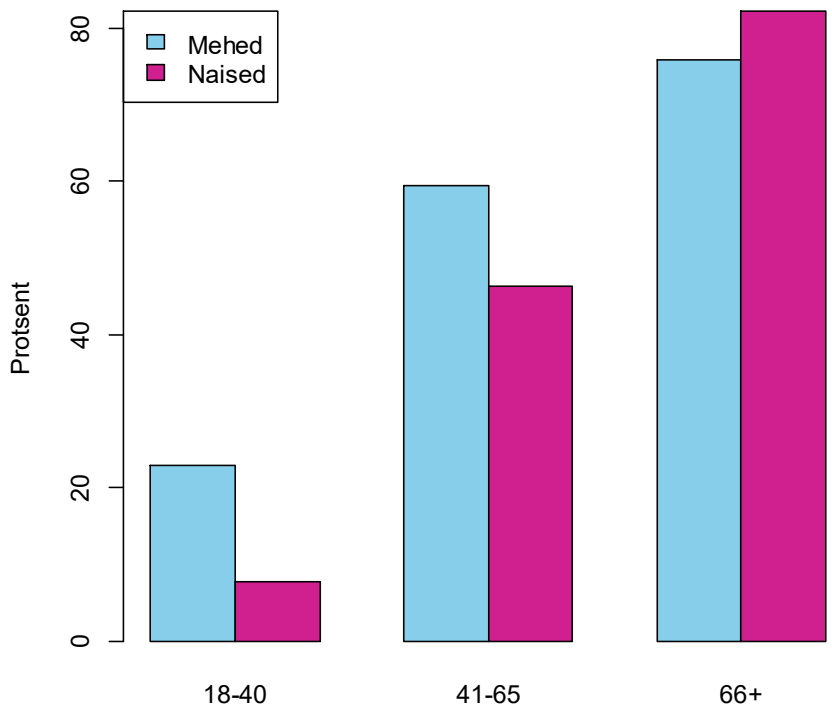
Suitsetamine: naised







Veel näiteid: Hüpertensiooni esinemine meestel ja naistel (TÜ EGV)



Gruppide võrdlemisel on hea kasutada lineaarset skaalat.

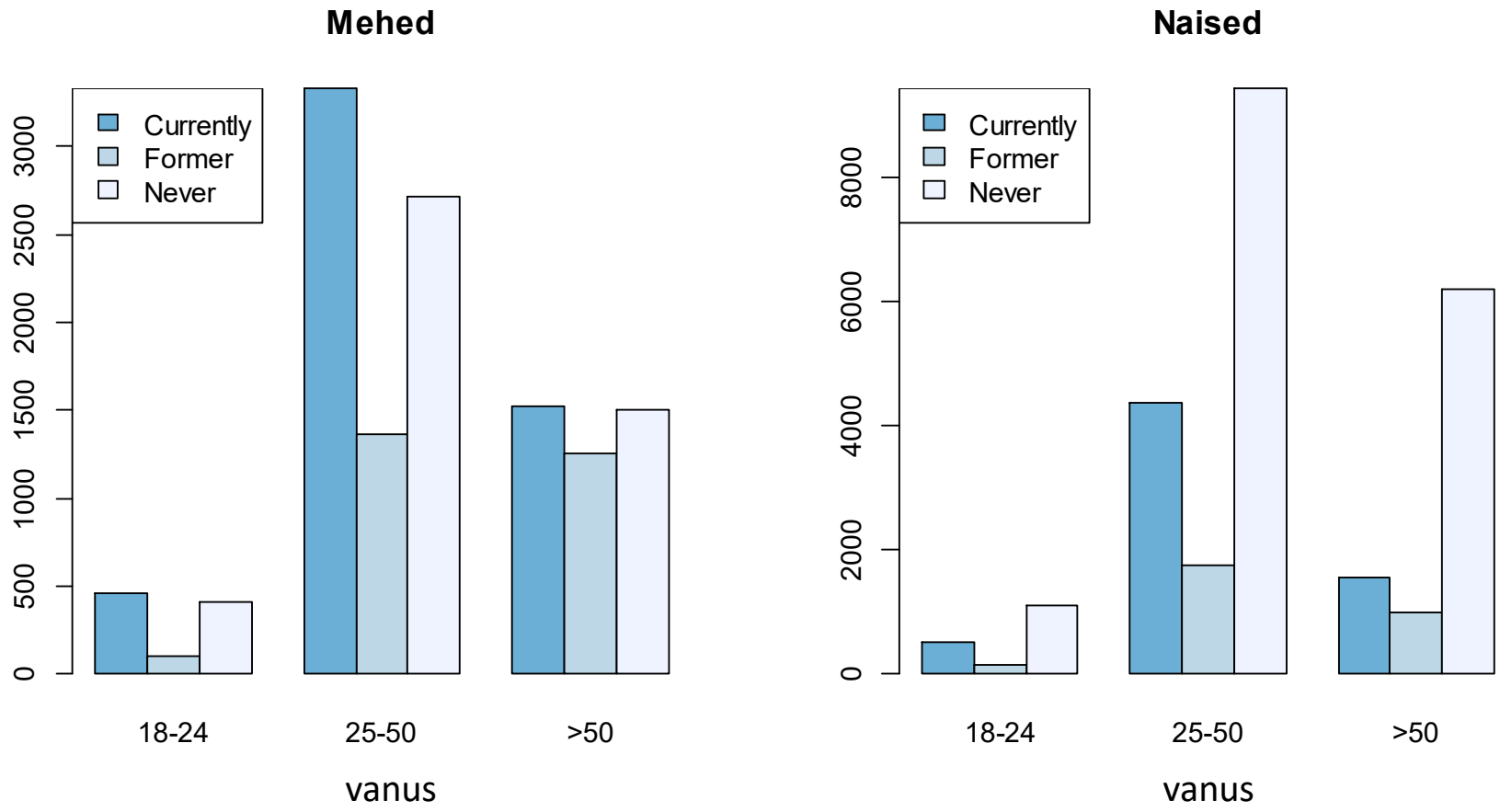
Tulpdiagrammi skaala algab 0-st!

Aga kui on vaja veel rohkem mõõtmeid?

	Vanus	Praegune suitsetaja	Endine suitsetaja	Pole suitsetanud	Kokku
Mehed	18-24	456	102	412	**970**
	25-50	3329	1361	2720	**7410**
	>50	1527	1254	1501	**4282**
Naised	18-24	517	145	1109	**1771**
	25-50	4371	1764	9440	**15575**
	>50	1545	991	6210	**8746**

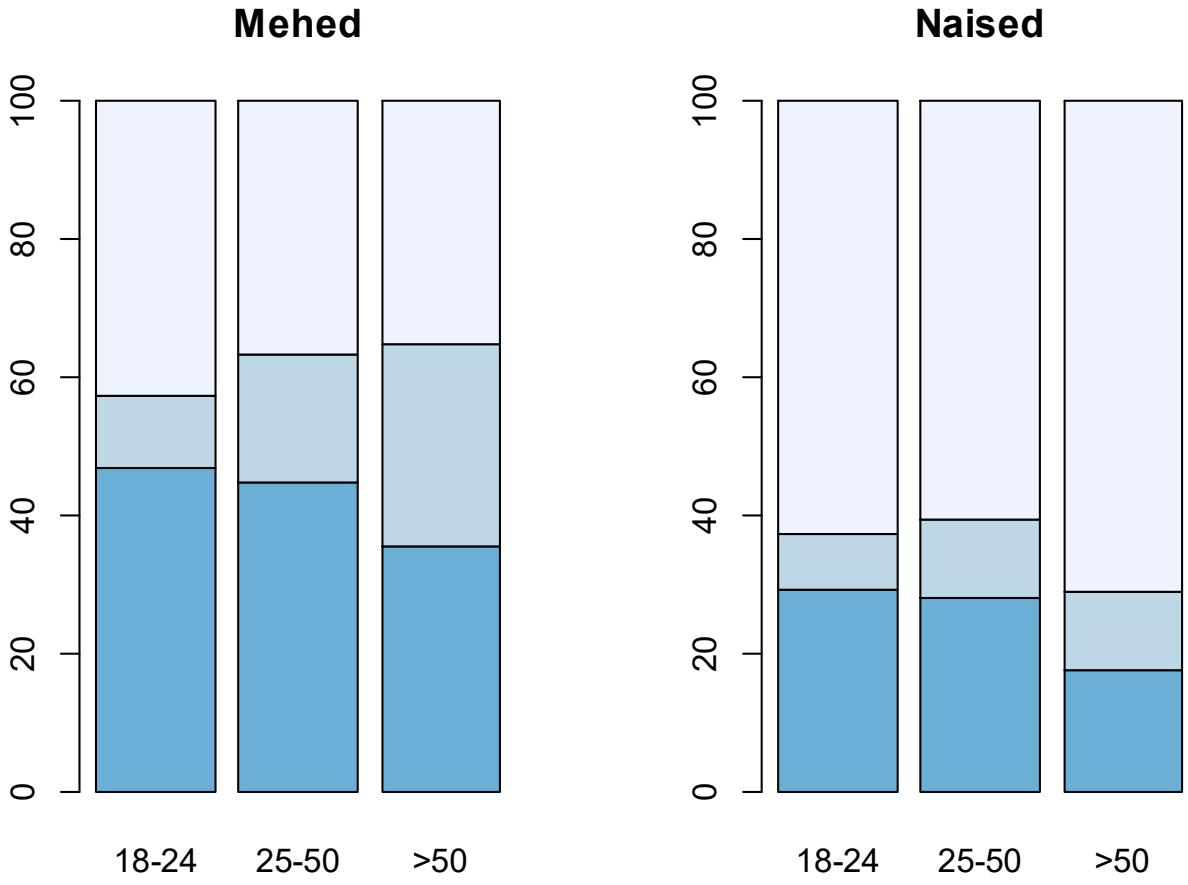
Millised protsendid võiks siia lisada?

Graafiline esitus?



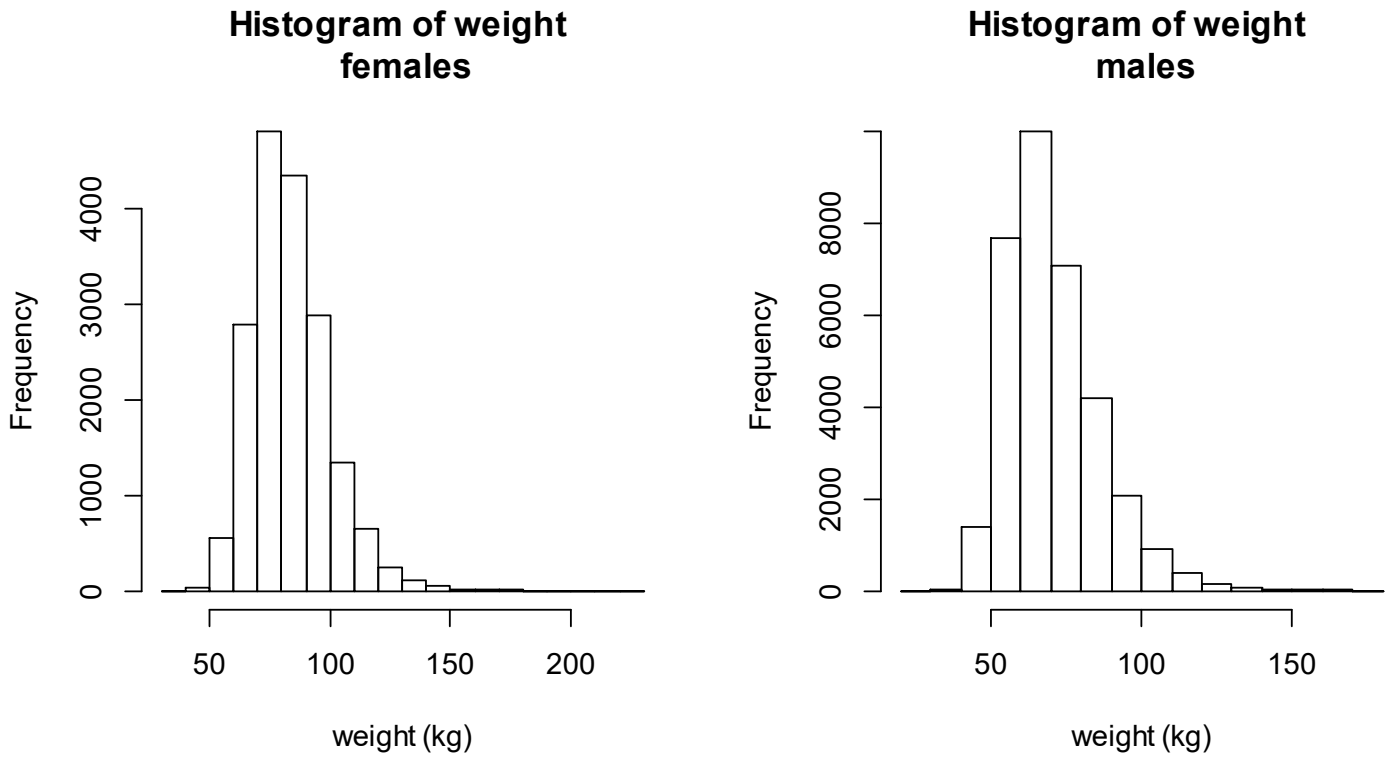
Graafiline esitus?

Ingl. stacked bar chart

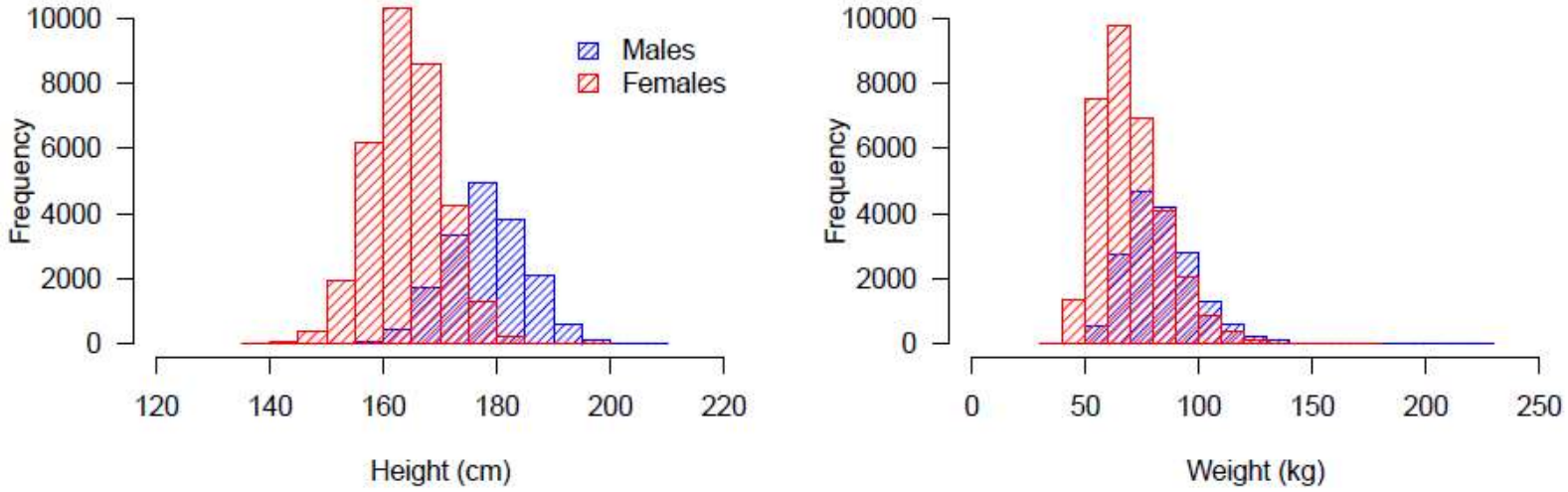


Pidevad tunnused

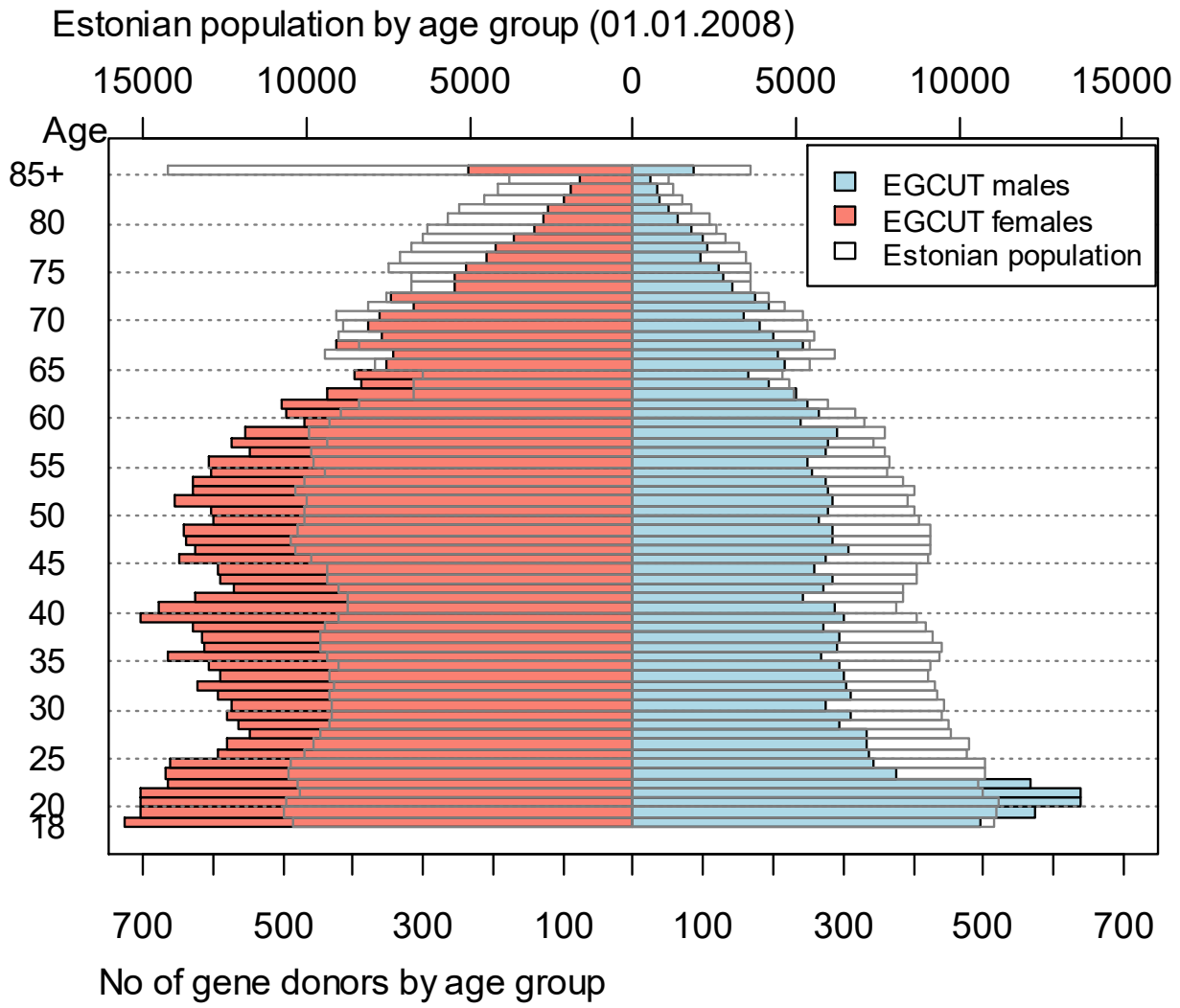
Pidevaid (tihti ka diskreetseid) tunnuseid iseloomustatakse tavaliselt kirjeldavate statistikute abil, nt keskmine *(mean)* ja standardhälve (*standard deviation, SD*). Graafikutest on tihti sobivaim histogramm.



Veidi keerulisemad histogrammid



.Ja veel….(populatsioonipüramiid)



Kirjeldavad statistikud

	Kaal (keskmine, standardhälve)	Pikkus (keskmine, standardhälve)
Mehed	84,4 (15,8)	178,5 (7,3)
Naised	71,3 (15,4)	164,7 (6,6)

Keskmine (keskväärtus):
Mean (average)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Vaatluste aritmeetiline keskmine

Standardhälve:
Standard deviation

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

Ruutjuur keskmisest ruuthälbest

Standardhälve (SD) iseloomustab **keskmist kaugust keskmisest**
ca 95% valimist asub vahemikus keskmine ± 2x SD

Mediaan

Kui andmed (arvulise tunnuse väärtused) sorteerida väiksemast suuremani, siis iga vaatluse järjekorranumbrit nimetatakse selle vaatluse astakuks (*rank*).

Näide

11 isiku pikkused (cm):

155, 160, 171, 182, 162, 153, 190, 167, 168, 165, 191

Järjestatud andmed:

153 155 160 162 165 **167** 168 170 171 182 191

Mediaan on järjestatud valimi keskpunkt – kas keskmise vaatluse väärtus (paarituarvulise valimisuuruse korral) või kahe keskmise vaatluse keskmine (paarisarvulise korral).

50% valimist on mediaanist suurem, 50% väiksem.

Miks eelistada keskmisele mediaani?

Näide 1

Ülesande lahendamiseks kulunud aeg (minutites) 15 õpilasel:

4 5 5 6 6 6 7 8 8 8 10 12 15 15 47

Keskmine on 10,8 minutit - paneme tähele, et sellest kauem kulus aega vaid 4 õpilasel ja see arv on mõjutatud selle poolt, et ühel õpilasel kulus erakordselt kaua aega

Leiame mediaani:

4 5 5 6 6 6 7 **8** 8 8 10 12 15 15 47

Mediaan on 8 minutit

Mediaan ei sõltu kõige aeglasema ega kõige kiirema õpilase tulemusest, vaid iseloomustabki nn keskmist osa õpilastest

Mediaani nimetatakse robustseks statistikuks, sest ta ei ole tundlik andmetes esinevate erandlikult suurte või väikeste väärtuste suhtes.

Aga vahel on siiski keskmine parem…

Näide 2

Hinded kahe grupi õpilastel:

Grupp 1:

2 2 2 3 3 3 3 **4** 4 4 4 4 4 4 5

Grupp 2:

3 3 4 4 4 4 4 **4** 5 5 5 5 5 5 5

Mediaan on mõlemas grupis sama: 4

Keskmine hinne on grupis 1 on 3,4 ning grupis 2 on see 4,3

Mediaani nimetatakse robustseks statistikuks, sest ta ei ole tundlik andmetes esinevate erandlikult suurte või väikeste väärtuste suhtes.