

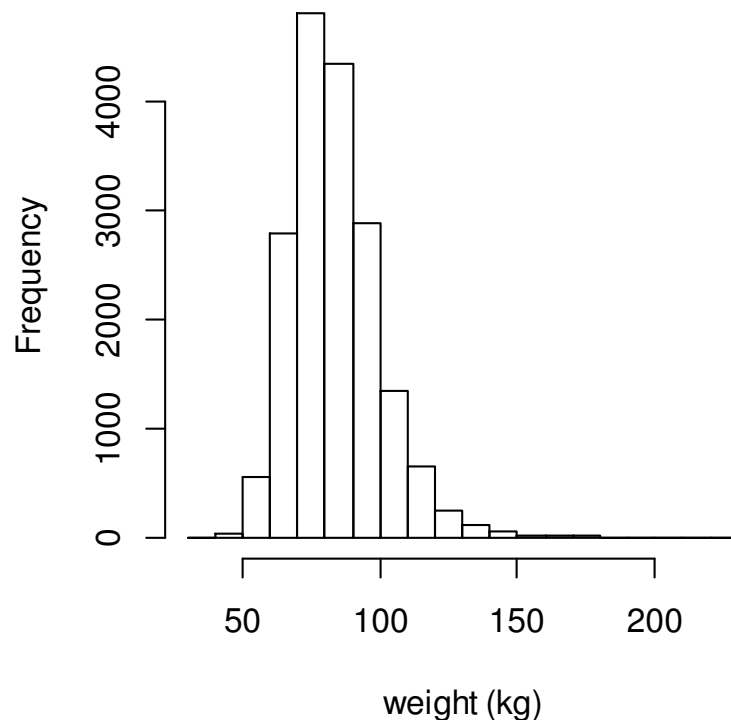
Kirjeldavad statistikud ja graafikud pidevatele tunnustele

Krista Fischer

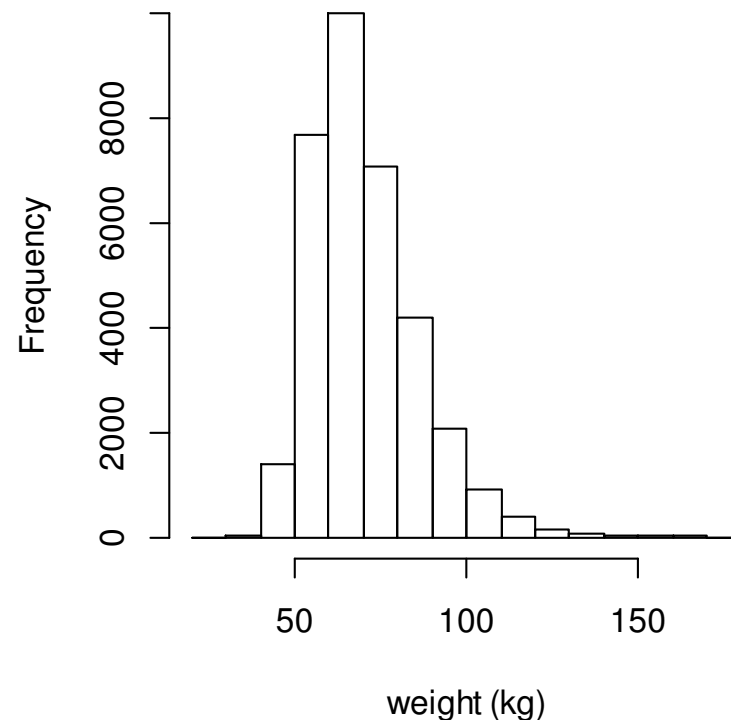
Pidevad tunnused ja nende kirjeldamine

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

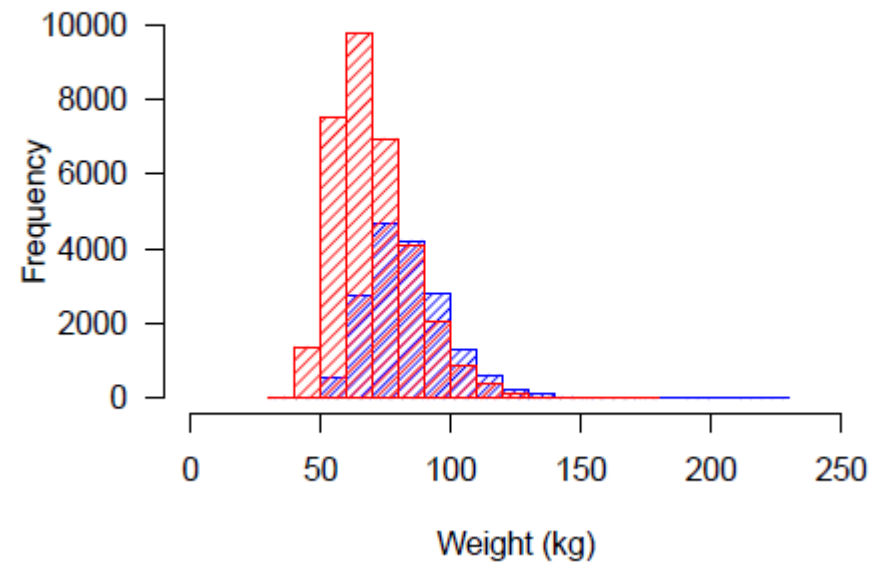
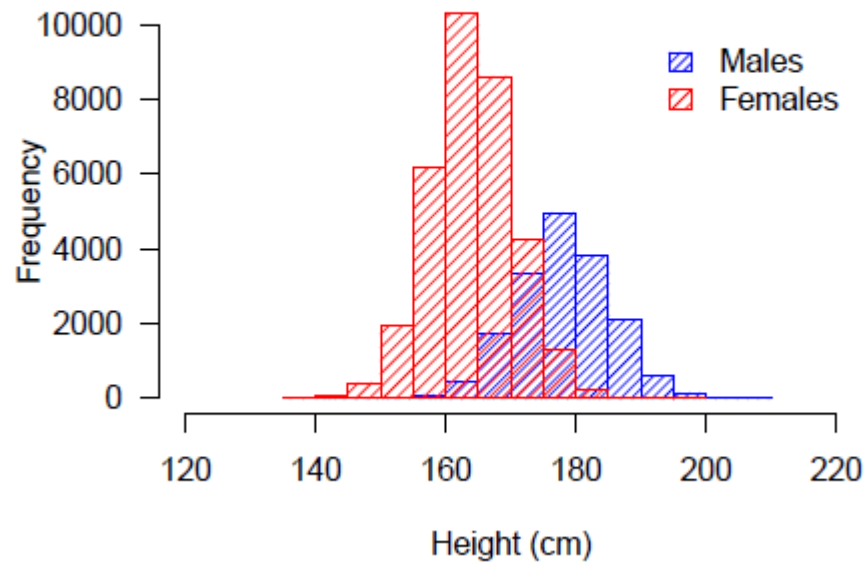
Histogram of weight
females



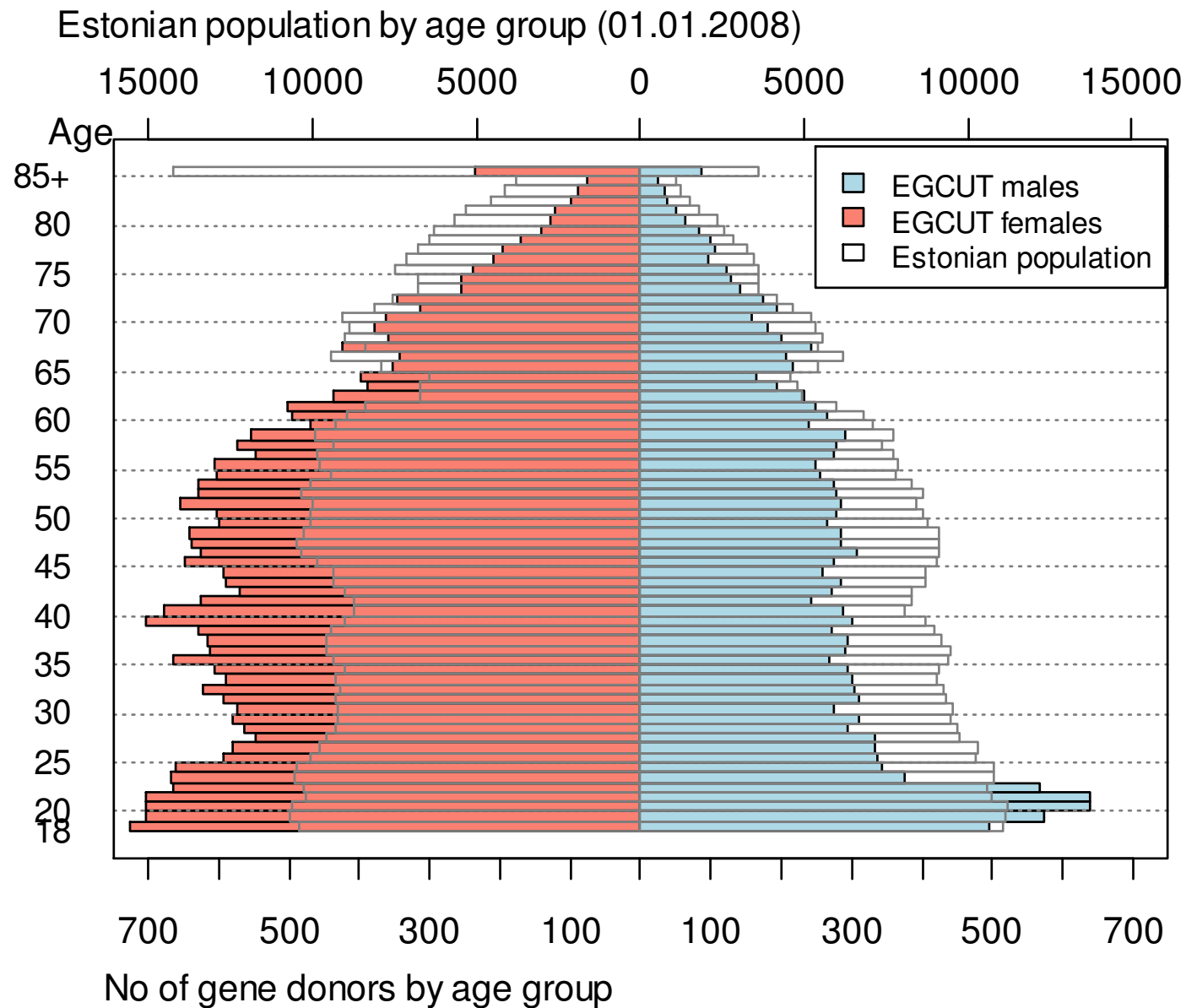
Histogram of weight
males



Veidi keerulisemad histogrammid



.Ja veel....(populatsioonipüramiid)



Kirjeldavad statistikud

	Weight (mean, SD)	Height (mean, SD)
Men	84,4 (15,8)	178,5 (7,3)
Women	71,3 (15,4)	164,7 (6,6)

Mean (average):
keskmine, keskväärtus

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

Standard deviation:
standardhälve

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

Standardhälve (SD) iseloomustab keskmist kaugust keskmisest –
95% valimist asub vahemikus keskmine $\pm 2 \times$ SD

Examples of data summaries:

Table 1. Baseline characteristics of the study participants.

	Estonian Biobank N=9,842	FINRISK 1997 N=7,503
Women — no. (%)	6,334 (64%)	3,741 (50%)
Age — years (range)	45.3 (18–103)	48.4 (24–74)
Body mass index (kg/m ²)	26.5±5.5	26.7±4.5
Systolic blood pressure (mmHg)	126±17	136±20
Current smokers — no. (%)	2,963 (30%)	1,770 (24%)
Use of antihypertensive therapy — no. (%)	2,489 (25%)	1,009 (13%)
Prevalent diabetes — no. (%)	737 (7.5%)	437 (5.8%)

(shortened)
Fischer et al. PlosMed 2014

Data are mean ± SD unless otherwise indicated.
[doi:10.1371/journal.pmed.1001606.t001](https://doi.org/10.1371/journal.pmed.1001606.t001)

Astakutel põhinevad statistikud: mediaan ja kvantiilid

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 (paaritu arvulise valimisuuruse korral) või kahe keskmise vaatluse keskmine (paarisarvulise korral).

50% valimist on mediaanist suurem, 50% väiksem.

Miks eelistada keskmisele mediaani?

- Näide

11 hiire eluiga päevades:

5 9 10 12 12 14 23 25 28 28 29

Ülejäänud 4 hiirt olid 30 päeva kestnud uuringu lõpus veel elus.
Kuidas neid andmeid esitada?

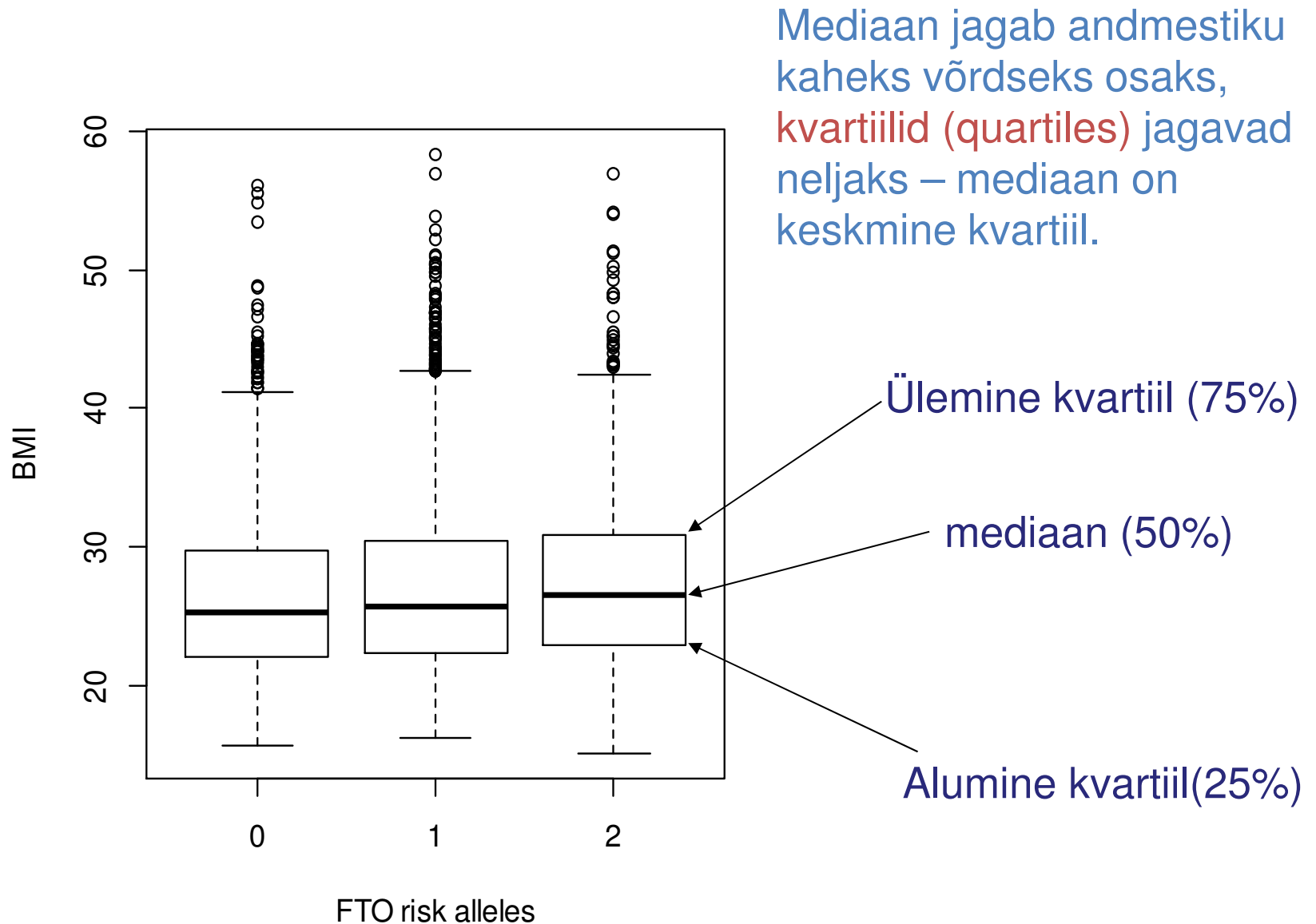
Keskmit ei saa arvutada, kuid andmed saab kirja panna nii:

5 9 10 12 12 14 23 25 28 28 29 30+ 30+ 30+ 30+

Elukestuse mediaan on 25 päeva

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

Karpdiagramm



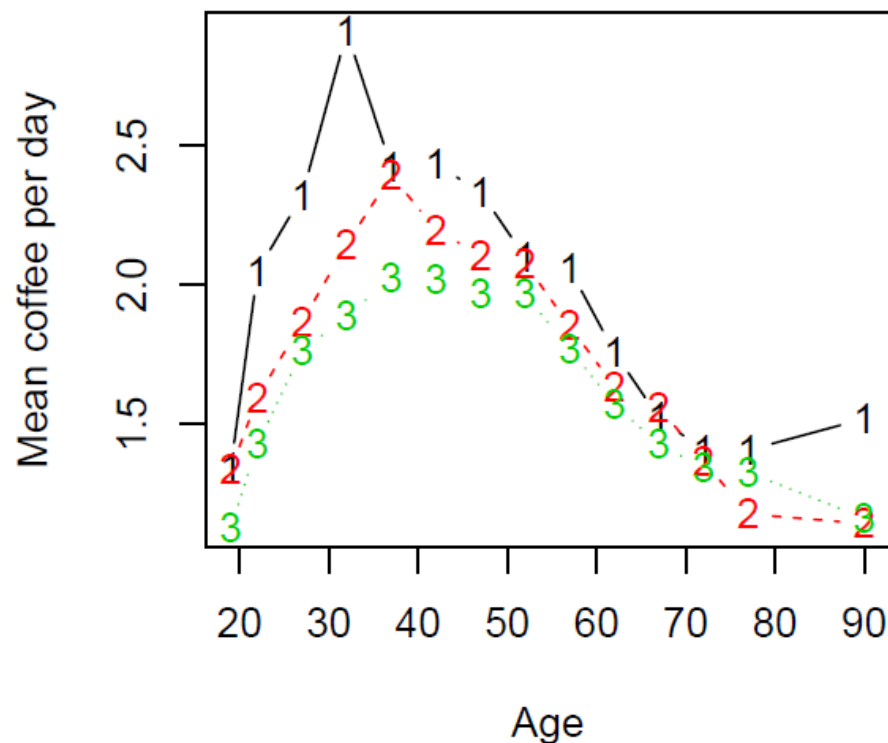
Aga näiteks siin on keskmine parem kui mediaan
(TÜ EGV: kohvi tarbimine vastavalt vanusele ja liitumise ajale):

Period 1: 2002-2005; Period 2: 2007-2008;

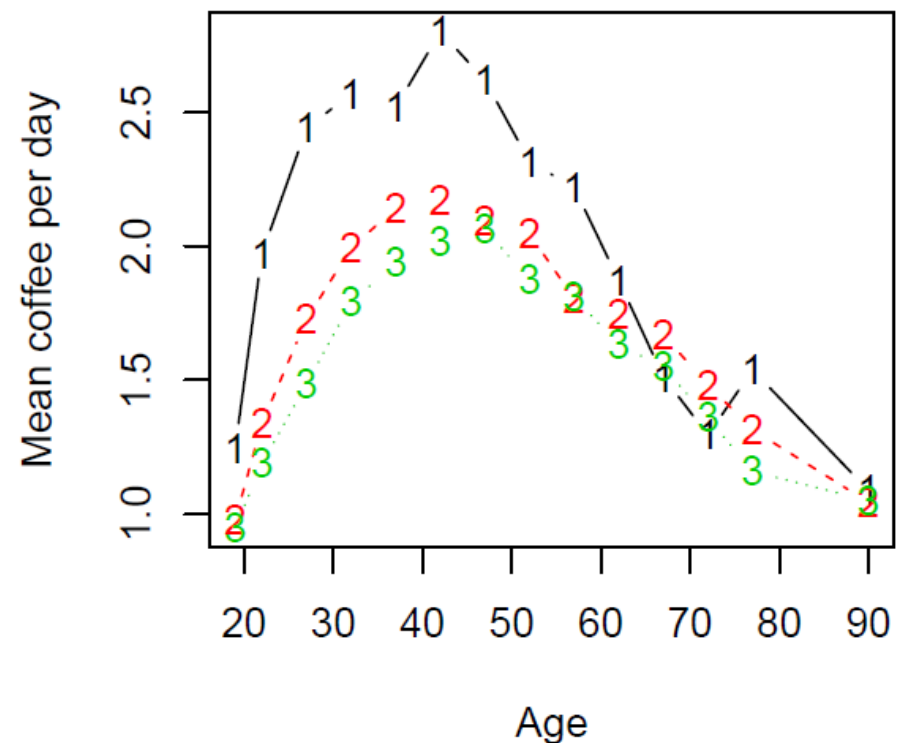
Period 3: 2009-2010

Cups of coffee per day

Males



Females

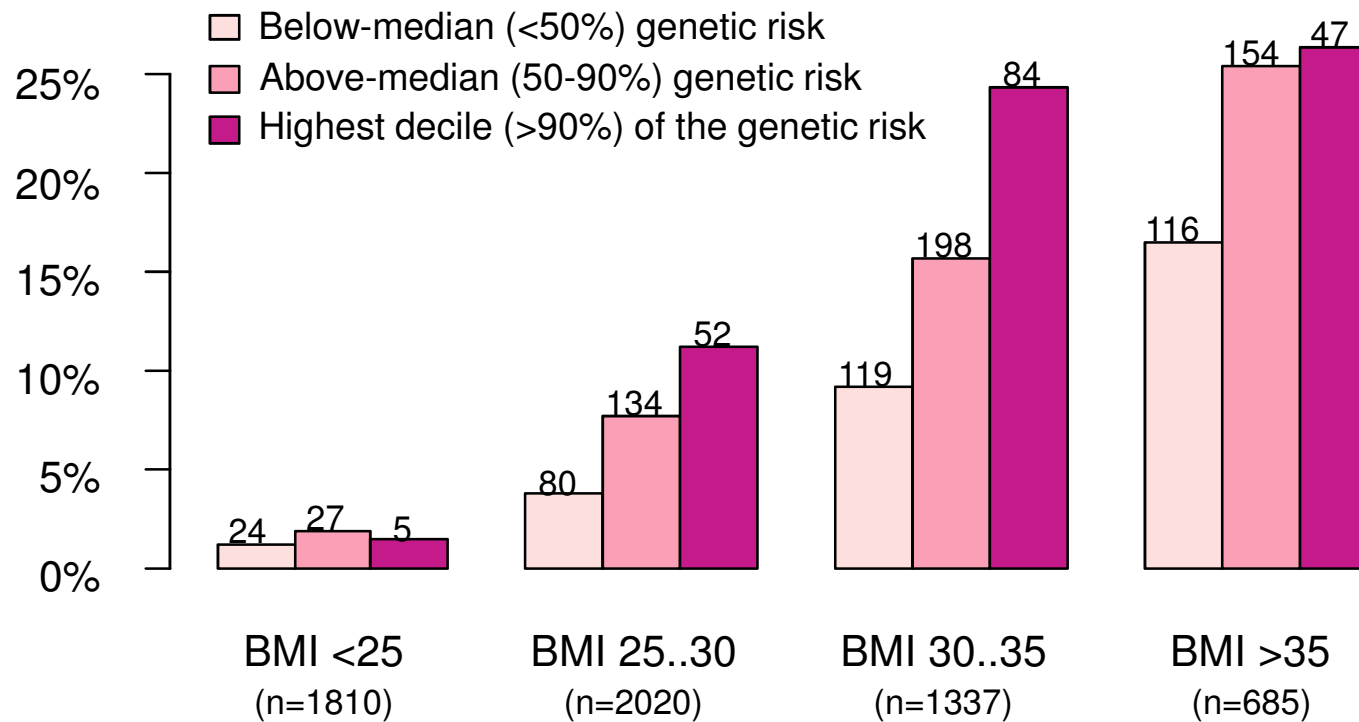


Teised kvantiilid (*quantiles*)

- $k\%$ kvantiil Q_k on defineeritud nii, et $k\%$ andmetest on väiksemad ja $(100-k)\%$ suuremad kui Q_k
- 10%, 20%,...,90% kvantiilid jagavad valimi 10 detsiiliks
- 20%, 40%, 60%, 80% kvantiilid jagavad valimi 5 kvintiiliks

Näide: geneetilise riskiskoori kvantiilid

T2D prevalence in individuals aged 45-80



Logaritmiline teisendus ja geomeetriline keskmine

- Kasutatakse siis, kui suhtelised erinevused on mõttekamad kui absoluutsed (väga tihti erinevate ainete kontsentratsiooni puhul mingites proovides, samuti geeniekspressiooni korral).

Nt väärtused 10, 10000, 300, 20, 1000

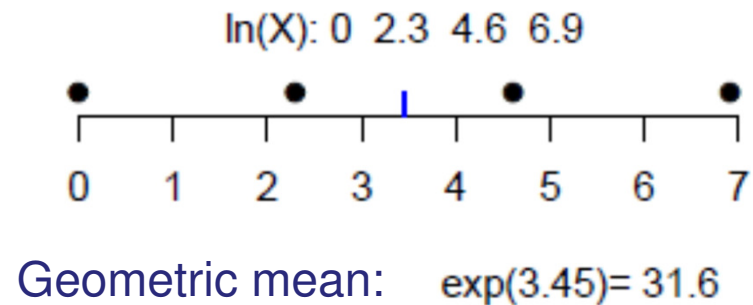
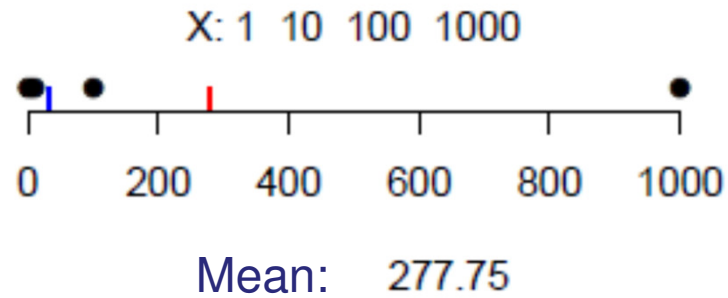
- Geomeetriline keskmine on defineeritud kui n -s juur vaatluste korrutisest – tavaliselt kasutatakse selle arvutamiseks logaritmimist

Geomeetrilise keskmise arvutamine:

- Leia logaritmitud tunnuse aritmeetiline keskmine
- Teisenda keskmine tagasi originaalskaalasse, kasutades astendamist (eksponenti)
- Oletame, et mingi aine kontsentratsioonid 4 inimese veres on 1, 10, 100 ja 1000. **Aritmeetiline keskmine on $1111/4 = 277.5$.**
- Kümnenlogaritmid ($\log_{10}$) neist väärtustest on 0, 1, 2 ja 3 ning nende keskmine $(0+1+2+3)/4 = 1.5$.

Geomeetriline keskmine on $10^{1.5} = 31.6$

Pole tähtsust, millist logaritmini kasutada – ka naturaallogaritm ($\ln$) annab sama tulemuse:



R funktsioonid

hist(x) - tunnuse x histogramm

mean(x) - keskmine

sd(x) - standardhälve

median(x) - mediaan

quantile(x, p= ...) - kvantiil, ... asemele arv 0-st 1-ni

boxplot(x) - karpdiagramm

boxplot(x~g) - karpdiagramm gruppide kaupa

Kokkuvõte

- Enne statistilise analüüsi juurde asumist, on kasulik andmeid visualiseerida ja kirjeldavaid statistikuid leida – lisaks esmase ülevaate saamisele saab nii leida ka andmetes esinevaid vigu
- Kirjeldavate kokkuvõtete tegemine andmetest võib olla sama aeganõudev kui keerukam statistiline analüüs