Sissejuhatus statistika erialasse

Statistilise analüüsi idee, statistilised seosed ja mudelid

Krista Fischer

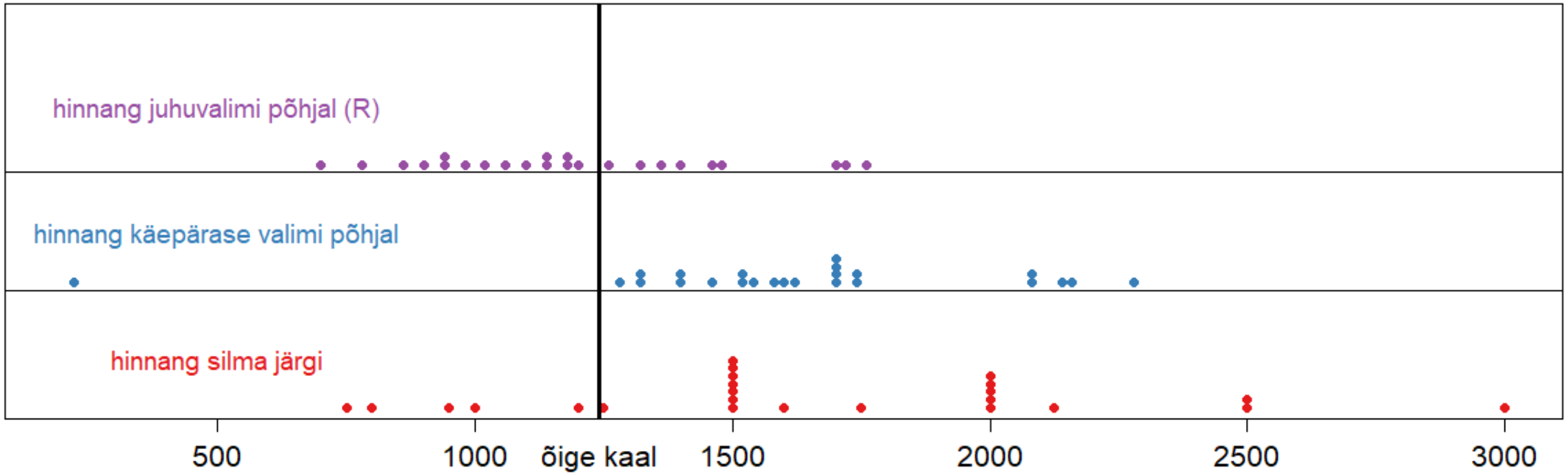
TÜ matemaatilise statistika professor

Tartu, 2019

I Statistilise analüüsi idee: usaldusvahemikud ja hüpoteeside testimine

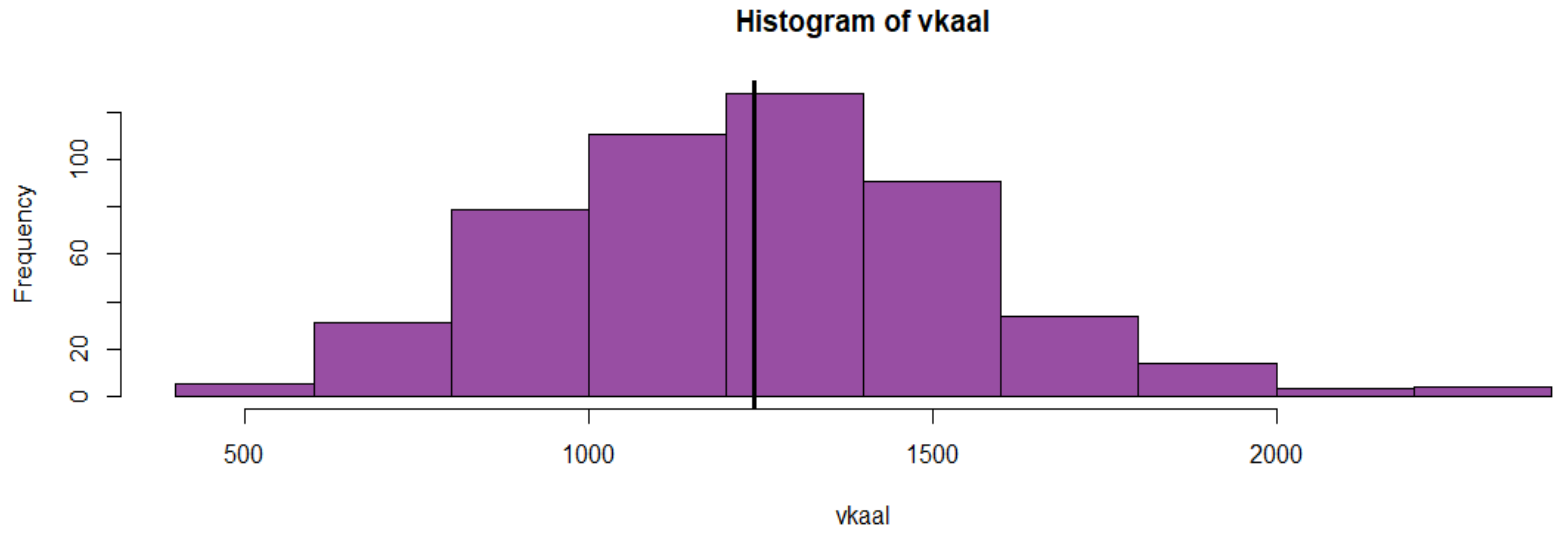
Meeldetuletus eelmisest loengust: valimite ja valiminäitajate varieeruvus

Hinnatud kommikoti kaal



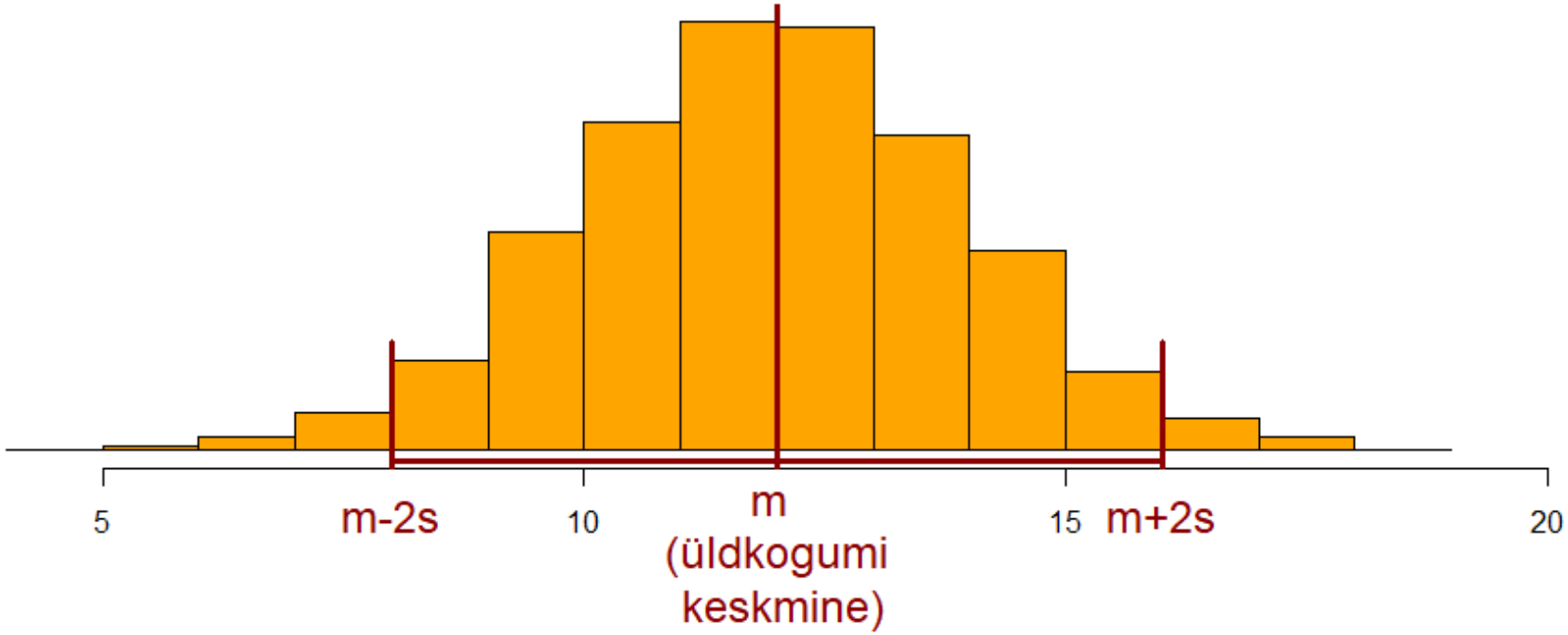
Valim on juhuslik, üldkogum mitte

Juhuvalimi põhjal arvutatud keskmised, protsendid jm näitajad on suure valimi korral ligikaudu normaaljaotusega, mille keskmine on üldkogumi vastav näitaja



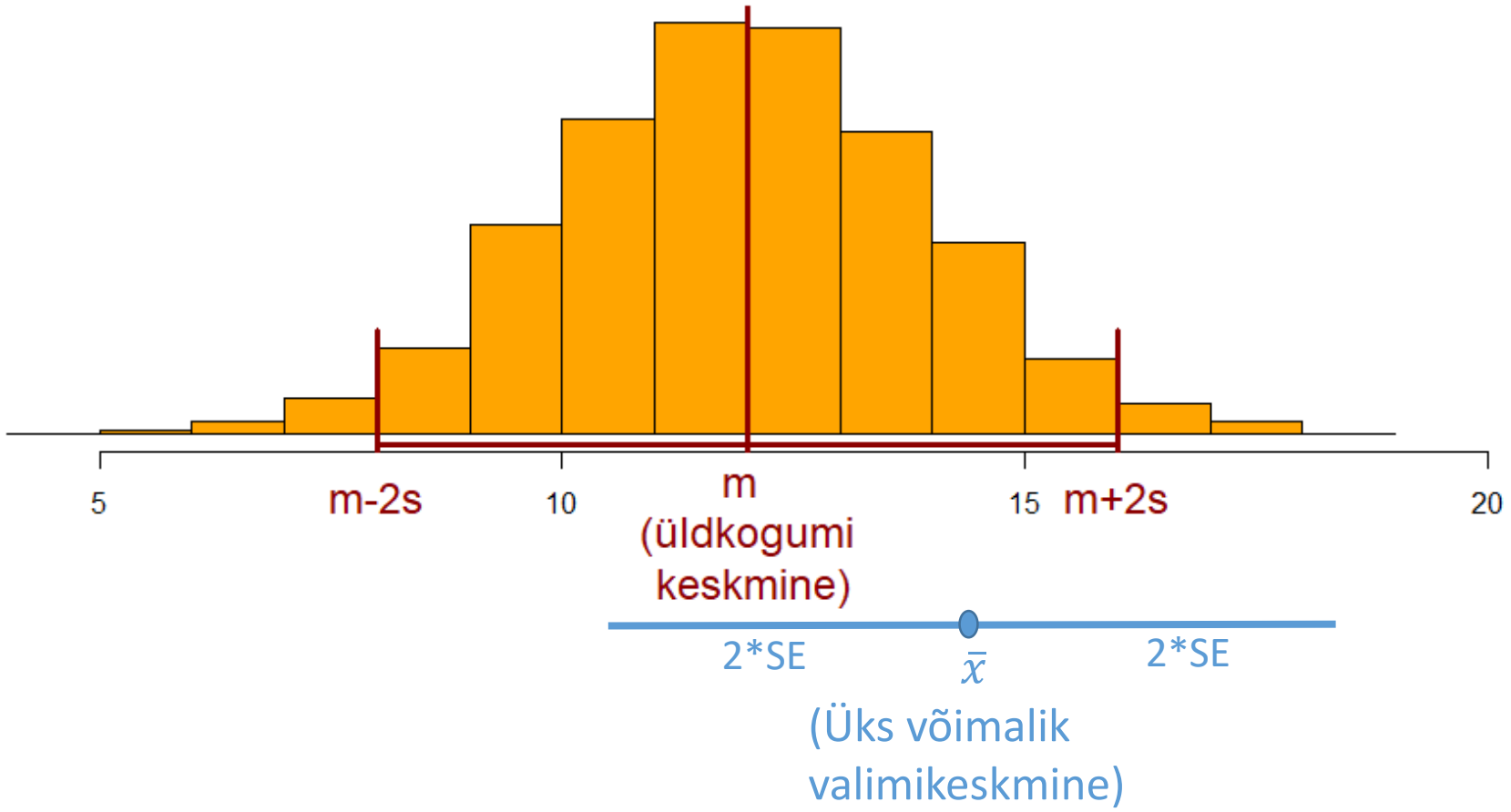
Kommide valim oli suhteliselt väike: n=20. Sellest ka väike ebasümmeetria (erinevus normaaljaotusest)

Normaaljaotusega suuruse puhul jääb 95% väärtustest vahemikku keskmine ± 2*standardhälve.



Seega on valimikeskmise puhul teada, et ta jääb 95% tõenäosusega mitte kaugemale kui 2 standardhälvet üldkogumi keskmisest.

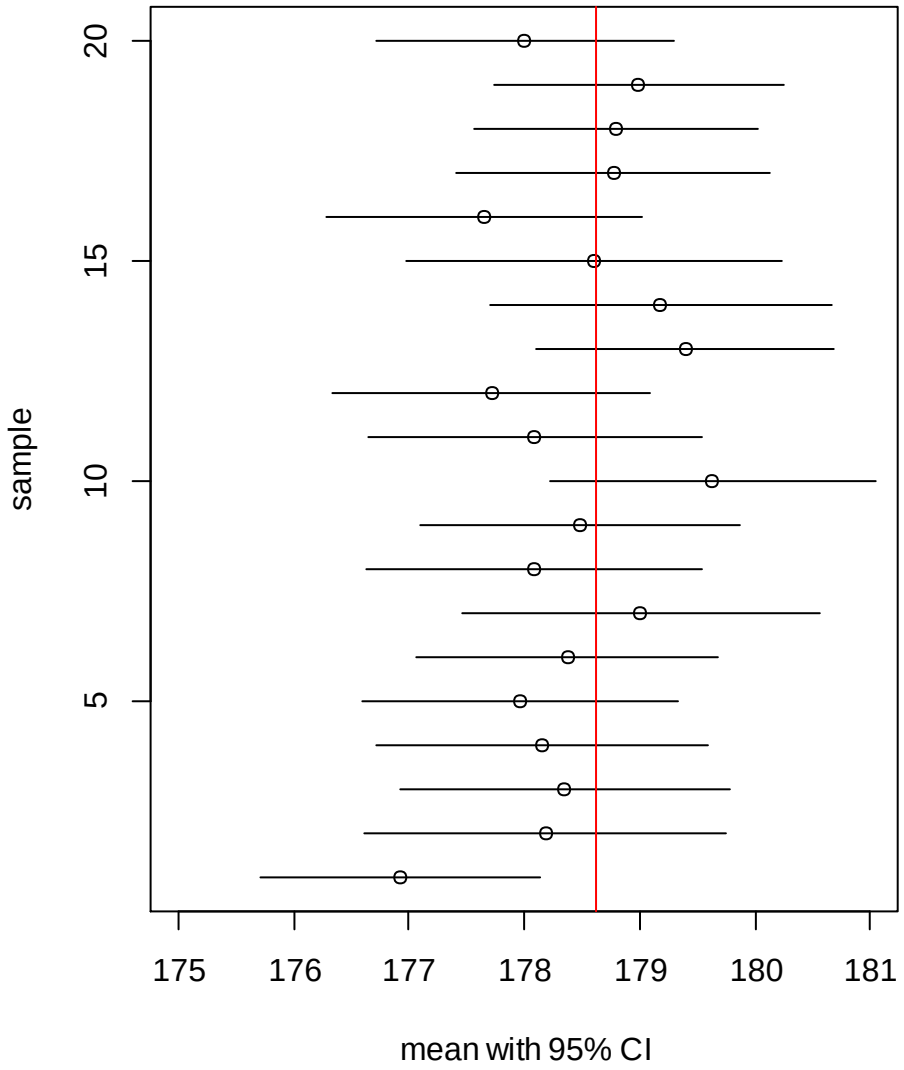
Valimikeskmise standardhälve kannab nimetust standardviga ja on valimi põhjal hinnatav kui $\mathbf{SE} = \frac{s}{\sqrt{n}}$, kus *s* on valimi standardhälve ja *n* valimi suurus



Tähistame valimikeskmist kui $\bar{x}$. Vahemik $\bar{x} \pm 2\mathrm{SE}$ kannab ligikaudse 95% usaldusintervalli ehk usaldusvahemiku nimetust

Usaldusintervallil on keskne roll statistilise analüüsi juures – enamasti arvutatakse see kõigile valimi põhjal saadud hinnangutele

Näide: 20 keskmist 95% usaldusvahemikuga



- Näide: TÜ Eesti Geenivaramu andmed – ainult need, kes on sündinud aastatel 1960-1974

Maakond	naised			mehed		
	N	Keskmine pikkus (cm)	95% usaldusintervall	N	Keskmine pikkus (cm)	95% usaldusintervall
Pärnumaa	358	167,9	167,3; 168,5	162	181,5	180,5; 182,5
Võrumaa	414	165,3	164,8; 165,9	173	179,2	178,2; 180,1
Tartumaa	885	166,6	166,2; 167,0	463	180,4	179,7; 181,0

Kas selle tabeli põhjal saab öelda, et …
- Pärnumaa naiste (antud vanusgrupi) keskmine pikkus on üle 167 cm?
- Pärnumaa meeste keskmine pikkus on alla 182cm?

Kas/miks ei piisa otsustamiseks vaid keskmistest pikkustest?

Kas selle tabeli põhjal saab järeldada, et Võrumaal elavad lühemad inimesed kui Pärnumaal?
Kas Tartumaal elavad pikemad inimesed kui Võrumaal, aga lühemad kui Pärnumaal?

Statistilised hüpoteesid

- Praktiliselt kõigi traditsioonilise statistilise analüüsi meetodite eesmärk on teatud statistiliste hüpoteeside kontroll.
- Tavaliselt on iga uuringu planeerimisel paika pandud tööhüpoteesid, mida soovitakse tõestada.

Näide. Soovitakse testida, et kõik Eestis elavad inimesed on lühemad kui 2m30cm. Kui mõõta üle 100, 1000 või ka 100000 eestimaalast ja need kõik jäävad lühemaks kui 2,30, siis ei ole see hüpotees siiski tõestatud (enne kui tõepoolest kõik on üle mõõdetud). Kui aga leitakse kasvõi üks inimene, kes on pikem kui 2m30cm, siis on see hüpotees kummutatud.

- Moraal: hüpoteesi kummutamine võib tihti olla praktiliselt teostatav ülesanne kui selle tõestamine
- Nullhüpotees (*null hypothesis*) on väide, mida peetakse tõeseks niikaua kuni andmetest ei leita piisavat tõendusmaterjali selle kummutamiseks – nullhüpoteesi ei tõestata. Kui nullhüpotees kummutatakse, siis see tähendab et võetakse vastu sisukas ehk alternatiivne hüpotees

Üks näide hüpoteeside kontrollimisest

- Oletame, et üks pood on otsustanud kampaania korras teha kingituse 20%-le ostjatest, kusjuures kingisaajad valitakse juhuslikult loosimise teel.
- Poe omanik avastab, et müüja on teinud kingituse 20-st esimest kliendist 7-le ja kahtlustab müüjat pettuses (jagas tuttavatele kingitusi?). Kas tal on õigus?

- Selge, et keskmiselt võiks 20-st inimesest 4 saada kingi, aga selge, et juhusliku varieeruvuse tõttu võib tegelik kingisaajate arv 20 inimese seas varieeruda.
- Kui suur varieeruvus on ootuspärane ja millest alates võib kahtlustada, et midagi on valesti?

Binoomjaotus

- Olgu X kingisaajate arv 20 inimese seas
- Tõenäosus, et mitte ükski inimene 20-st ei saa kingitust, on: $P(X = 0) = 0{,}8^{20} = 0{,}012$.
- Tõenäosus, et täpselt üks inimene saab kingituse, on $P(X = 1) = 20 \cdot 0{,}2 \cdot 0{,}8^{19}$. (20 võimalikku kombinatsiooni, igaühe tõenäosus on $0{,}2 \cdot 0{,}8^{19}$).
- Tõenäosus, et k inimest saavad kingituse, on:

$$P(X = k) = C_{20}^{k}\ 0{,}2^{k}(1 - 0{,}2)^{20-k},$$

  kus C_{20}^{k} on kombinatsioonide arv 20-st k kaupa.

- Saame öelda, et X on binoomjaotusega, parameetritega 20 ja $0{,}2$. Üldine valem:

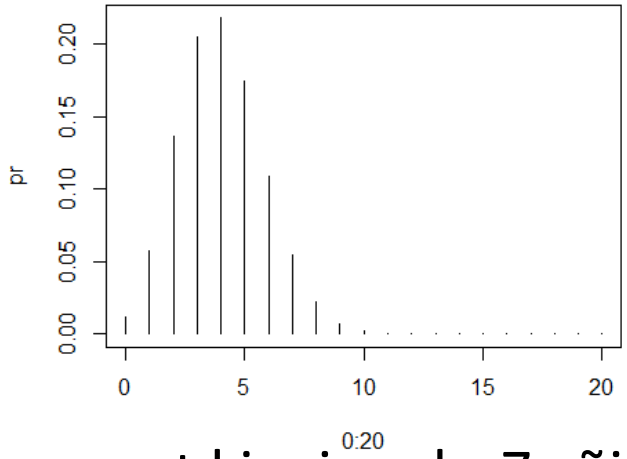
$$P(X = k) = C_{n}^{k}\ p^{k}(1 - p)^{n-k},$$

  kus C_{n}^{k} on kombinatsioonide arv n-st k kaupa

Seega...

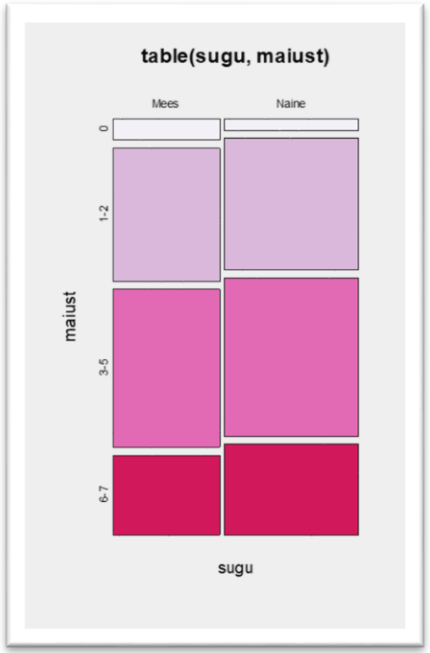
- Kingisaajate hulk 20 kliendi seas on binoomjaotusega, parameetritega 20 ja 0,2. Vaatame, millise tõenäosusega on võimalik saada kink 0…20 kliendil. Samuti hindame tõenäosust, et kingi saab 7 või rohkem inimest 20-st:

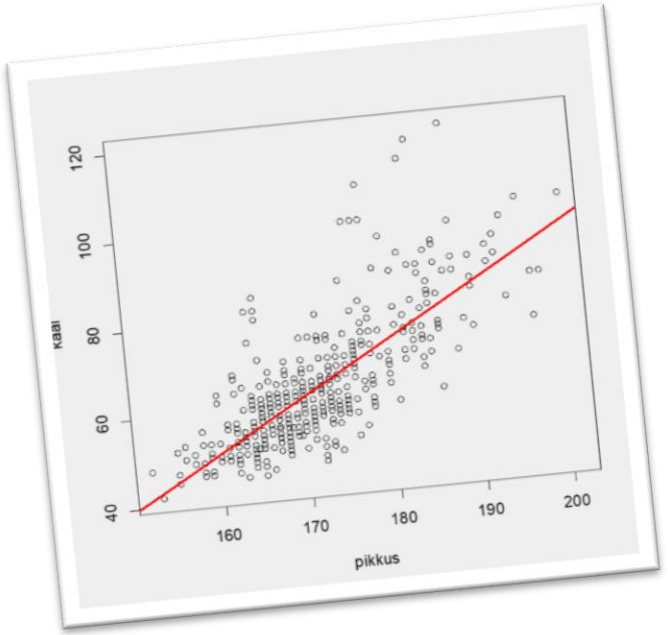
```
R: pr<-dbinom(0:20,20,0.2)
   plot(pr,type="h")
   1-pbinom(6,20,0.2)
```

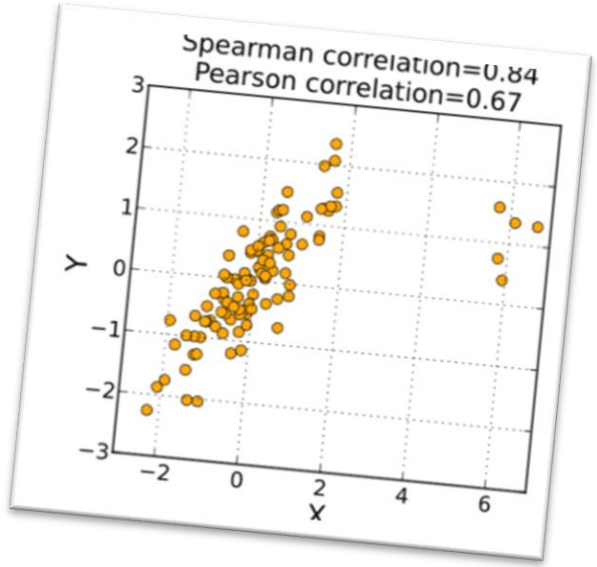


- Näeme, et ausa loosimise korral on tõenäosus, et kingi saaks 7 või enam inimest, 8,7%
- Kas see tõenäosus on suur või väike?
- See tõenäosus (tõenäosus, et nullhüpoteesi kehtimisel saadakse vaadeldud või äärmuslikum väärtus) kannab p-väärtuse ehk olulisuse tõenäosuse nimetust. Kokkuleppeliselt kummutatakse nullhüpotees, kui p<0,05 (ehk 5%)
- Seda reeglit järgides saame otsustada, et andmed on pigem kooskõlas hüpoteesiga, et loosimine toimus ausalt

II Statistilised seosed







Statistilised seosed

- Meeldetuletus: sõltumatud vs sõltuvad sündmused
- P(õhtul sajab vihma) või

  P(õhtul sajab vihma|hommikul on taevas pilvine) või

  P(õhtul sajab vihma| kuu faasiks on „täiskuu“)

- Sündmused A ja B on sõltumatud, kui sündmuse ühe toimumise tõenäosus ei sõltu teise sündmuse toimumisest
- Sõltumatute sündmuste korral: P(A ja B) = P(A)P(B)

Tihti on statistilise analüüsi eesmärgiks hinnata, kas tunnused on sõltumatud või esineb neil omavaheline seos

- Seos suitsetamise ja infarkti riski vahel?
- Seos poliitilise erakonna eelistuse ja hariduse vahel?
- Seos poes kõlava muusika ja ostueelistuste vahel?
- Seos aktsiahindade ja USA valimistulemuste vahel?
- Seos pikkuse ja kaalu vahel?
- Seos hariduse ja kehamassiindeksi vahel?
- Seos kehamassindeksi ja eluea vahel?
- Seos kohvijoomise ja eluea vahel?

Üks lihtne reegel: alusta alati seose graafilisest uurimisest!

- Kuidas luua graafik, mis toob esile võimaliku seose olemasolu või puudumise?
- Graafilise meetodi (ja hiljem ka analüüsimeetodi) valik sõltub tunnuste tüüpidest

Näide: kas värskete puuviljade söömine sõltub soost?

- Värskete puuviljade söömisel on oluline seos mitmete haiguste riskiga (risk on väiksem neil, kes rohkem vitamiinirikkaid toite tarbivad)
- Eestis on meeste keskmine eluiga väga palju lühem kui naistel (74 vs 82 aastat).
- Kas osaliselt on see selgitatav sellega, et meeste eluviisid kipuvad olema ebatervislikumad?

R-i abil saadud toortabel (geenivaramu andmed):

```
> table(f1$vpuuv,f1$sugu)

      M     N
1  1060  1219       # ei söö üldse
2  5918  7204       # 1-2 päeval nädalas
3  5716 11584       # 3-5 päeval nädalas
4  3627 12392       # 6-7 päeval nädalas
```

Teeme veidi huvitavama tabeli (veeruprotsentidega):

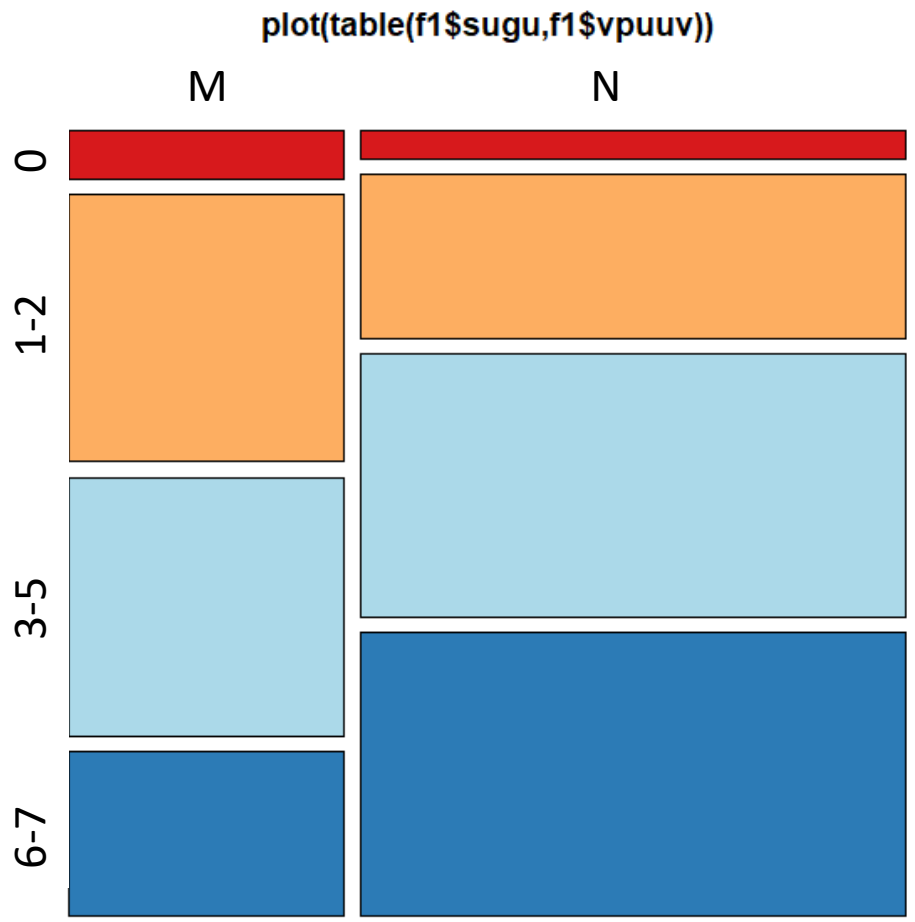
```
> stat.table(list(vpuuv,sugu),contents=list(count(),percent(vpuuv)),data=f1)
 ------------------------
        ------sugu-------
vpuuv           M        N
 ------------------------
1             1060     1219  arv
               6.5      3.8  protsent

2             5918     7204
             36.3     22.2

3             5716    11584
             35.0     35.8

4             3627    12392
             22.2     38.2
 ------------------------
```

Näeme, et 6,5% meestest ja 3,8% naistest ei söö üldse puuvilju.

Igapäevaselt sööb puuvilju 22,2% meestest ja 38,2% naistest

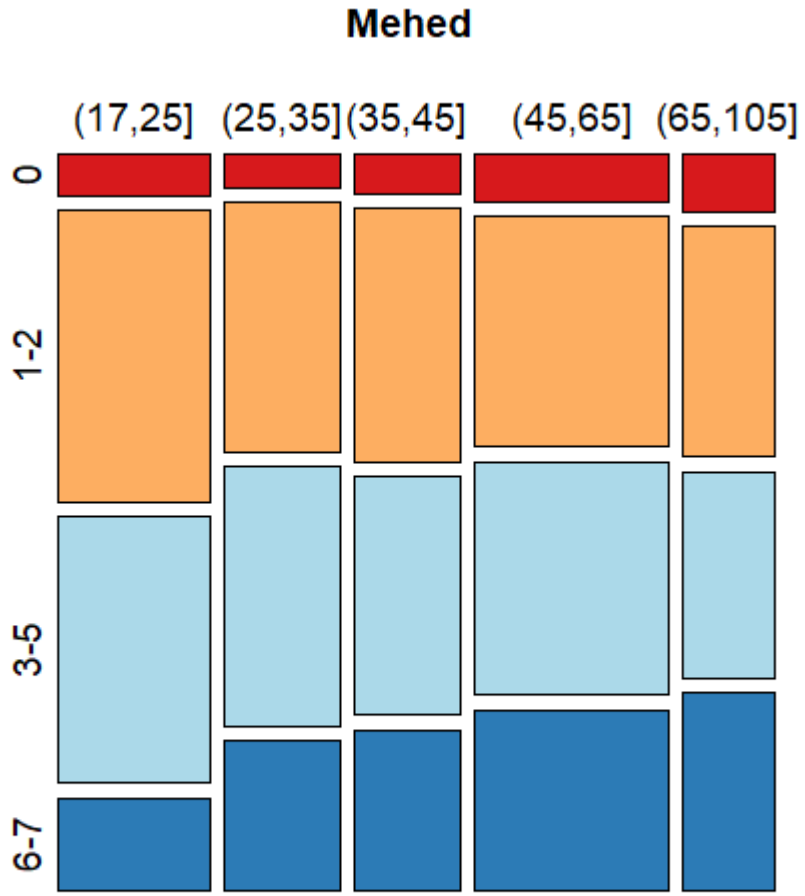
Kuidas seda graafiliselt kujutada?

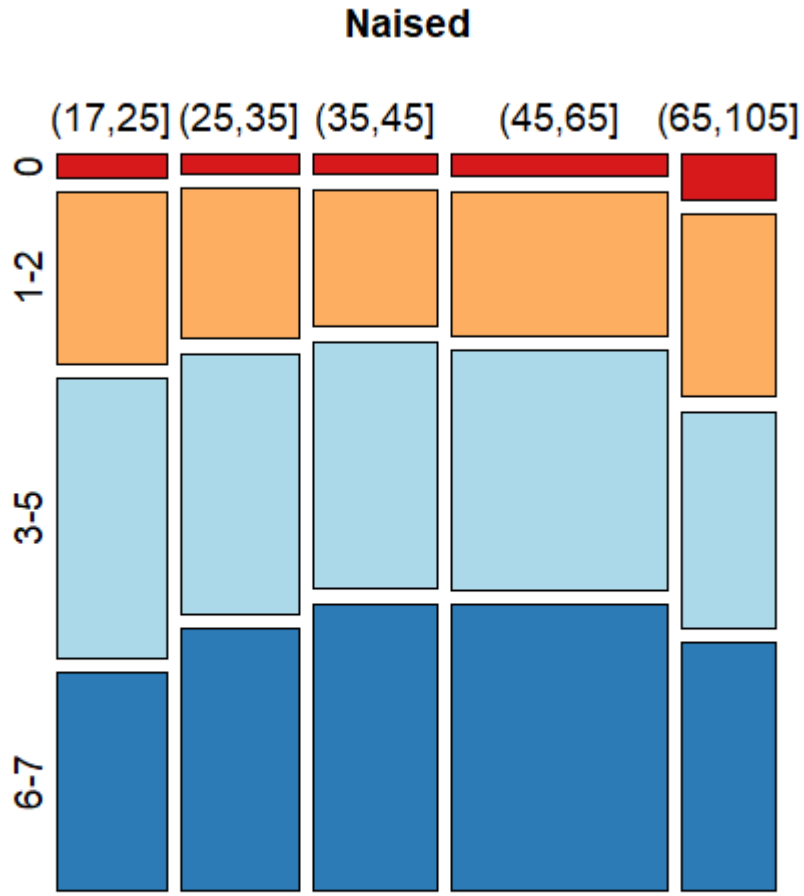
Üks võimalik joonis:



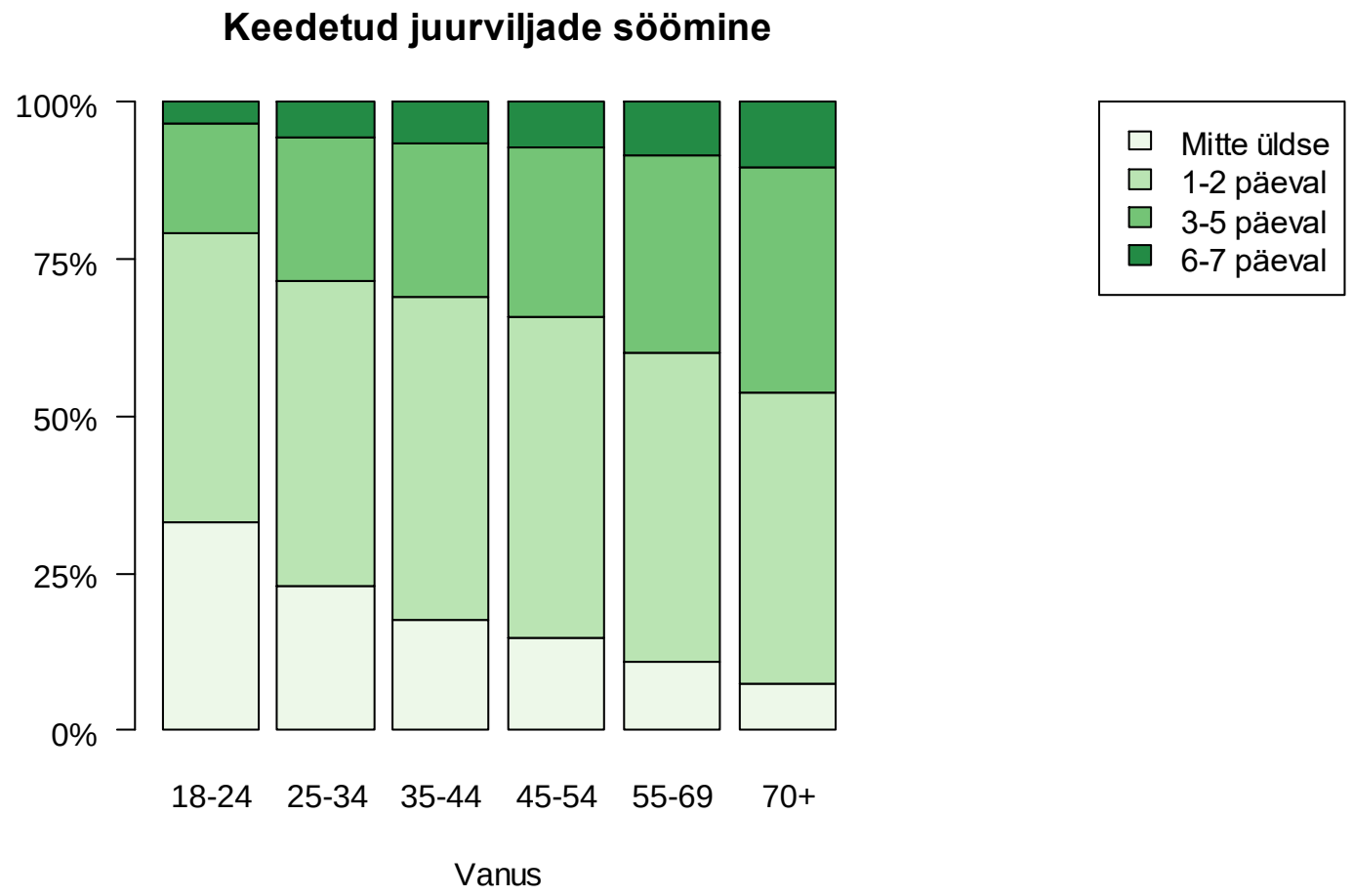
Nn *Mosaic plot:* tulba laius vastab grupi suurusele (proportionaalselt), värviliste tulpade kõrgused protsentuaalsele jaotusele

Lisame ka vanuse



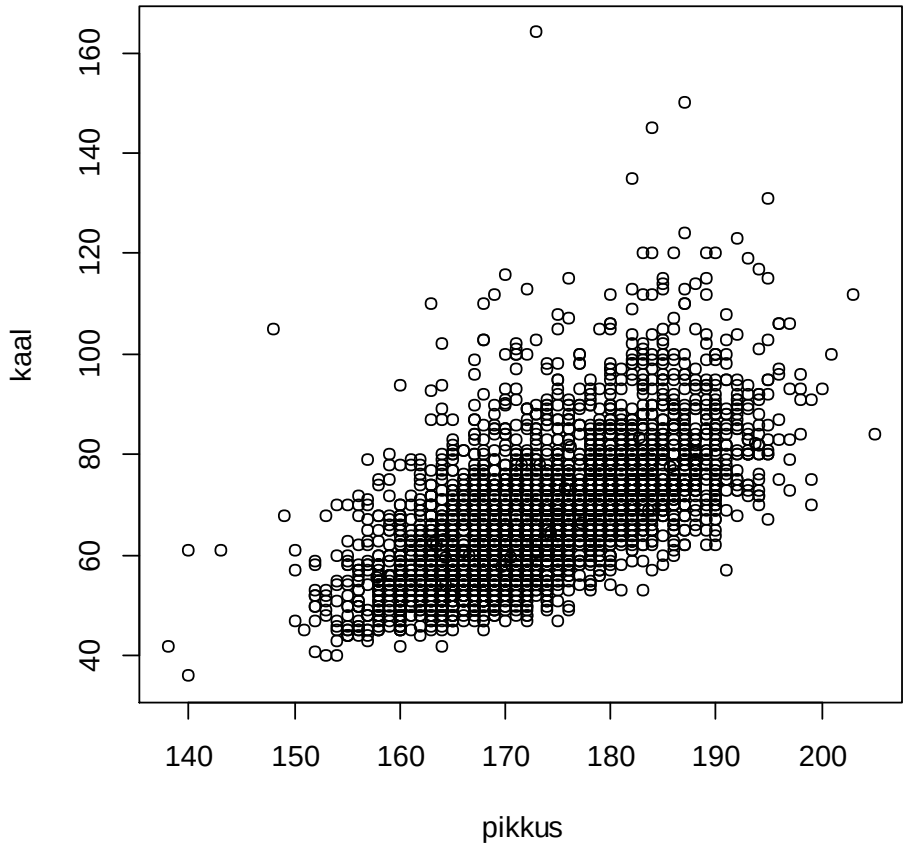


Veidi lihtsam joonis (aga kaob ära grupisuuruste info):

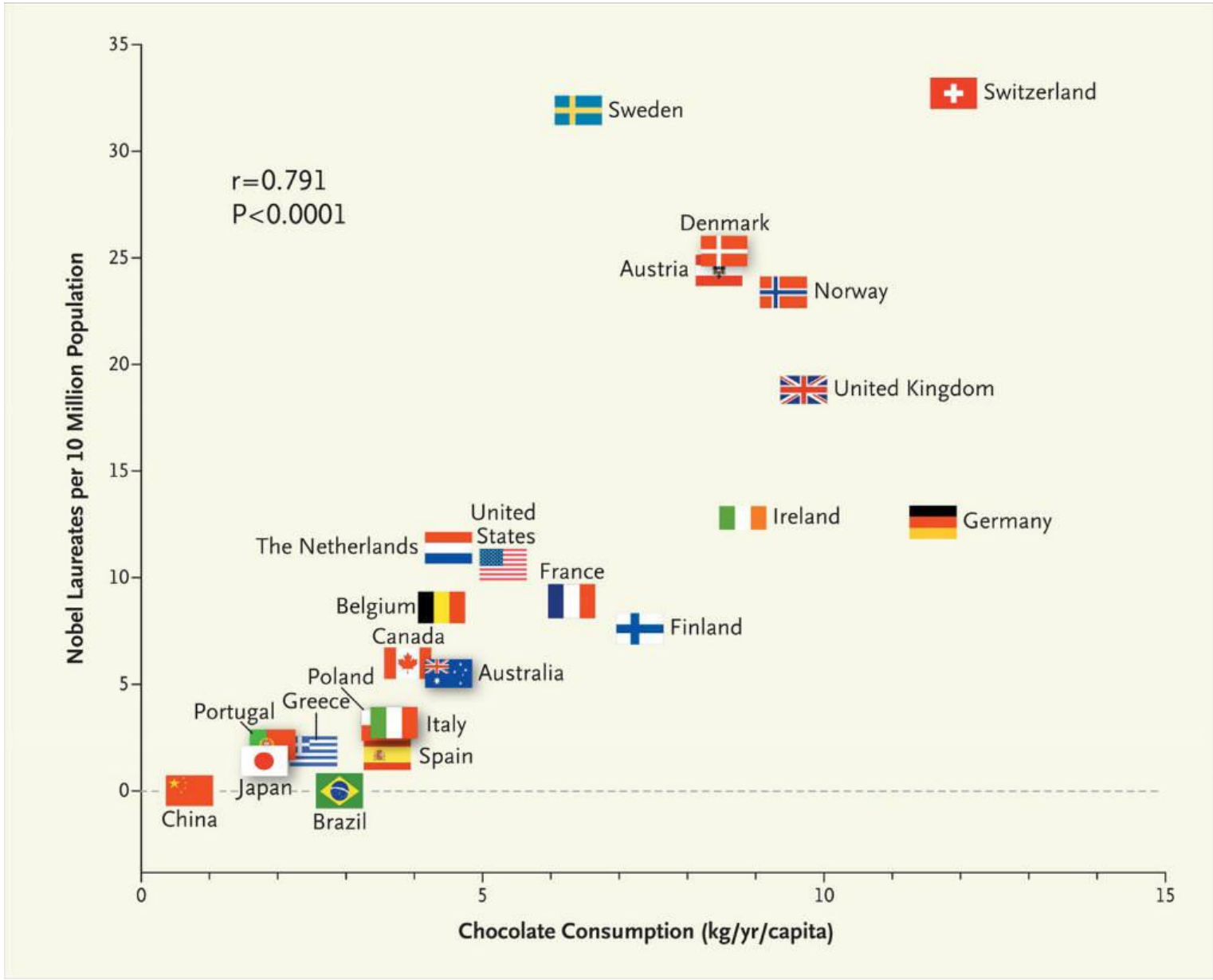


Kaks pidevat tunnust

Lihtne hajuvusdiagramm *(scatter plot)*



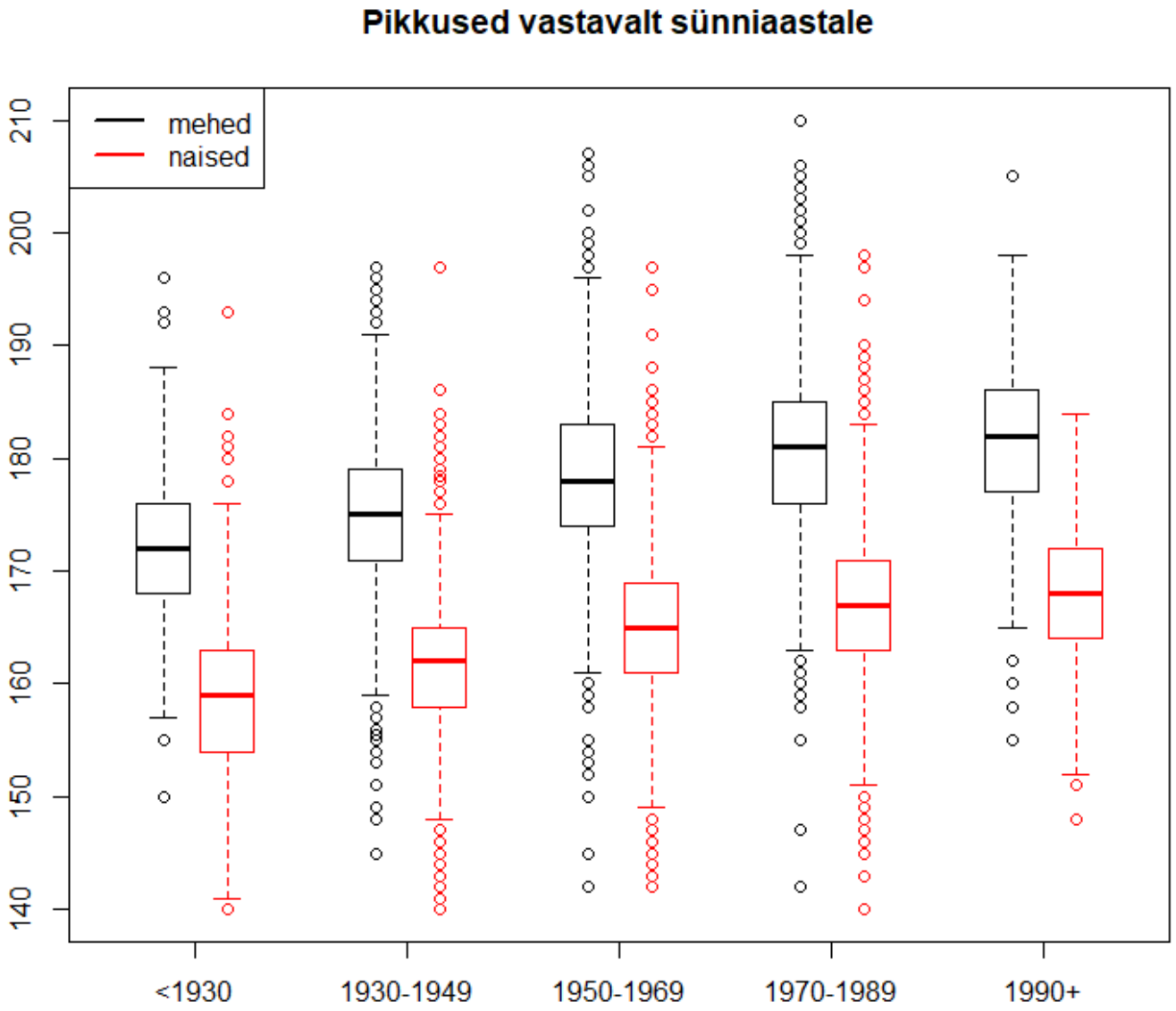
R: plot(x,y)



https://www.nejm.org/doi/full/10.1056/NEJMon1211064

(Messerli, F., *New England Journal of Medicine*, 2012)

Pidev tunnus ja nominaalne- või järjestustunnus: grupeeritud karpdiagramm (Box plot)



Kahe tunnuse vahelised seosed erinevate andmetüüpide korral

	Pidev tunnus	**Binaarne tunnus**	**Nominaalne tunnus (categorical variable)**
Pidev utnnus (Continuous variable)	Hajuvusdiagramm Korrelatsioon Lineaarne regressioon		
Binaarne tunnus	Grupeeritud karpdiagramm, T-test, Wilcoxoni test, logistiline regressioon	2x2 tabel, šansisuhe, riskisuhe	
Nominaalne tunnus	Grupeeritud karpdiagramm, error bar plot, dispersioonanalüüs	2xk tabel, grupeeritud tulpdiagramm hii-ruut test	m x k tabel, grupeeritud tulpdiagramm hii-ruut test

Seose uurimine ja testimine sagedustabelis: Hii-ruut testi idee

Näide: maiustuste tarbimine geenivaramu andmebaasis 18-20-aastaste noorte hulgas

	Ei söö	1-2 päeval nädalas	3-5 päeval nädalas	6-7 päeval nädalas
Mehed	94	589	696	353
Naised	64	735	881	510

Kas maiustuste tarbimise sagedus sõltub soost?

Seose uurimine ja testimine sagedustabelis

Näide: maiustuste tarbimine geenivaramu andmebaasis 18-20-aastaste noorte hulgas

	Ei söö	1-2 päeval nädalas	3-5 päeval nädalas	6-7 päeval nädalas	Kokku
Mehed	94	589	696	353	**1732**
Naised	64	735	881	510	**2190**
Kokku	**158**	**1324**	**1577**	**863**	**3922**

Kas maiustuste tarbimise sagedus sõltub soost?

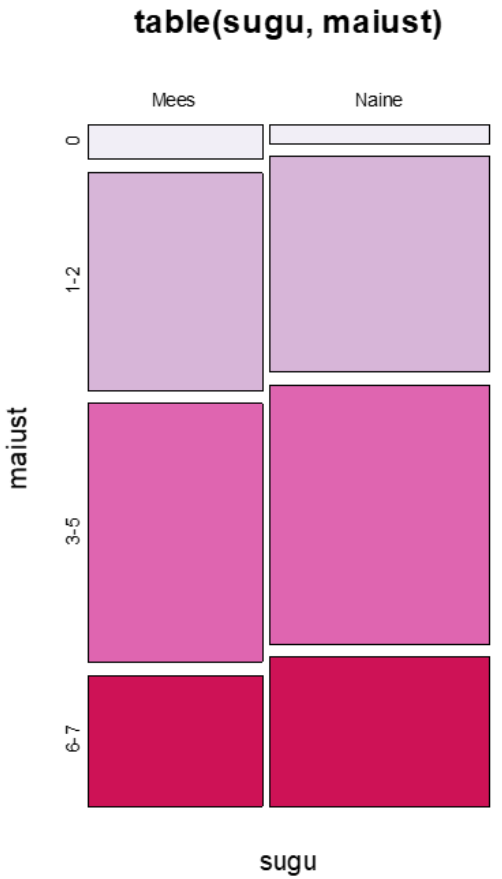
Seose uurimine ja testimine sagedustabelis

Näide: maiustuste tarbimine geenivaramu andmebaasis 18-20-aastaste noorte hulgas

Arvutame reaprotsendid

	Ei söö	1-2 päeval nädalas	3-5 päeval nädalas	6-7 päeval nädalas	Kokku
Mehed	94 (5.4%)	589 (34.0%)	696 (40.2%)	353 (20.4%)	**1732 (100%)**
Naised	64 (2.9%)	735 (33.6%)	881 (40.2%)	510 (23.3%)	**2190 (100%)**
Kokku	**158 (4.0%)**	**1324 (33.8%)**	**1577 (40.2%)**	**863 (22.0%)**	**3922**

Kas maiustuste tarbimise sagedus sõltub soost?



Seos soo ja maiustuste söömise vahel

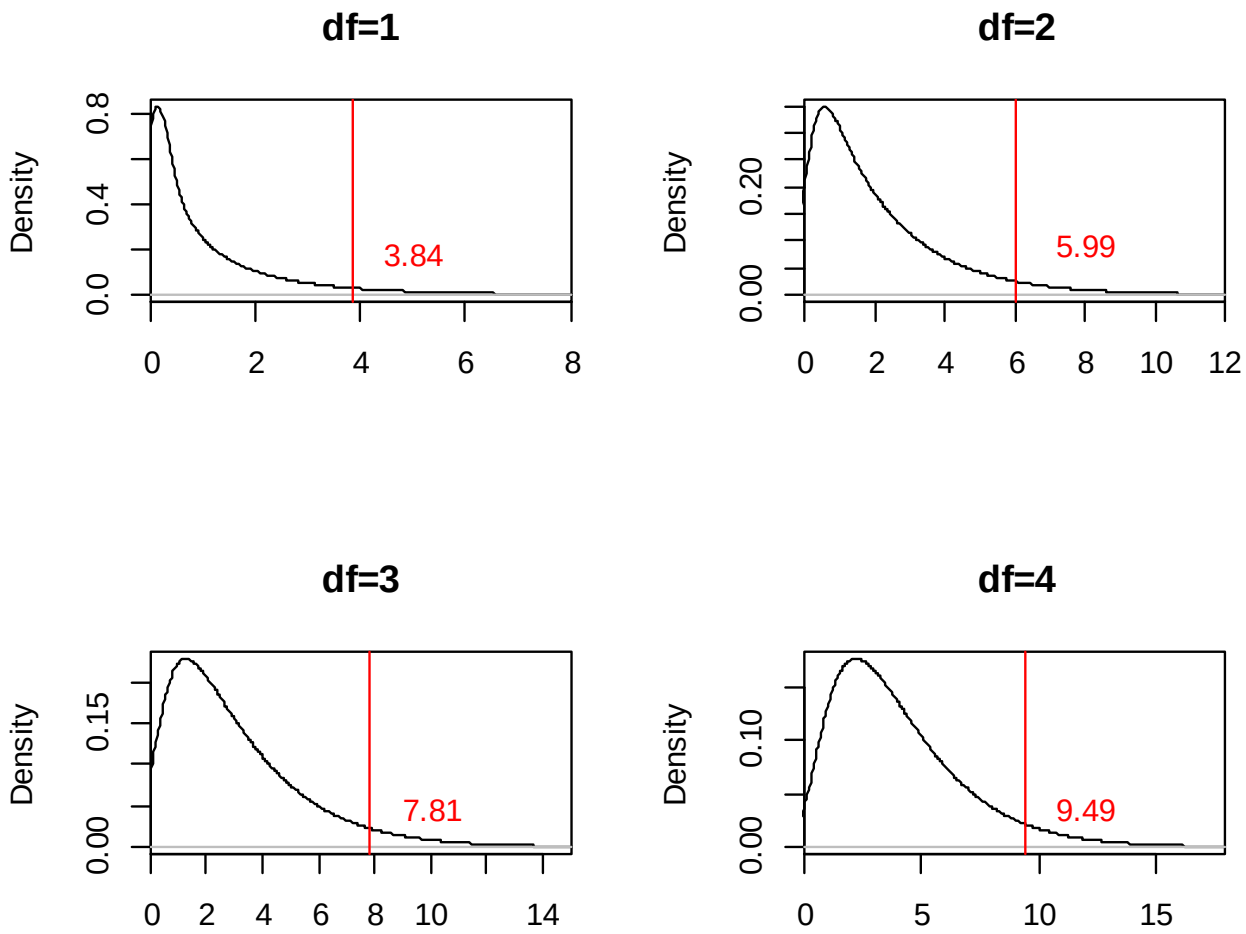
	Ei söö	1-2 päeval nädalas	3-5 päeval nädalas	6-7 päeval nädalas	Kokku
Mehed	94 (5.4%) 70 (4.0%)	589(34.0%) 585 (33.8%)	696 (40.2%) 696 (40.2%)	353 (20.4%) 381 (22.0%)	**1732 (100%)**
Naised	64 (2.9%) 88 (4.0%)	735(33.6%) 739 (33.8%)	881 (40.2%) 881 (40.2%)	510 (23.3%) 482 (22.0%)	**2190 (100%)**
Kokku	**158 (4.0%)**	**1324 (33.8%)**	**1577 (40.2%)**	**863 (22.0%)**	**3922**

Hii-ruut testi (χ^2 test) idee: võrdleme tegelikke ja oodatavaid sagedusi
Kui O_{ij} on vaadeldud (*observed*) sagedus i-ndas reas ja j-ndas veerus ning E_{ij} vastav oodatud (*expected*) sagedus, siis arvutatakse χ^2 statistik järgmiselt:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{v}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad \text{kus } \chi^2 \sim \chi^2(df) \text{ ja } df = (r-1)\times(v-1)$$

Siin r on ridade arv ja v veergude arv tabelis; df - χ^2 statistiku vabadusastmete arv (määrab statistiku jaotuse nullhüpoteesi kehtimisel)

Hii-ruut statistiku jaotused nullhüpoteesi kehtimisel (χ^2 –jaotuse tihedusfunktsioonid) erineva vabadusastmete arvu korral ja 95% kvantiil (väärtus, millest suurema statistiku väärtuse korral kummutatakse nullhüpotees, st $p<0.05$)



	Ei söö	1-2 päeval nädalas	3-5 päeval nädalas	6-7 päeval nädalas	Kokku
Mehed	94 (5.4%) 70 (4.0%)	589(34.0%) 585 (33.8%)	696 (40.2%) 696 (40.2%)	353 (20.4%) 381 (22.0%)	**1732 (100%)**
Naised	64 (2.9%) 88 (4.0%)	735(33.6%) 739 (33.8%)	881 (40.2%) 881 (40.2%)	510 (23.3%) 482 (22.0%)	**2190 (100%)**
Kokku	**158 (4.0%)**	**1324 (33.8%)**	**1577 (40.2%)**	**863 (22.0%)**	**3922**

Analüüs R-i abil:

```
> chisq.test(table(sugu,maiust))

        Pearson's Chi-squared test

data:  table(sugu, maiust)
X-squared = 18.833, df = 3, p-value = 0.000296
```

III Statistiliste mudelite maailm

Mille poolest erinevad ja mille poolest sarnanevad automudel ja päris auto?

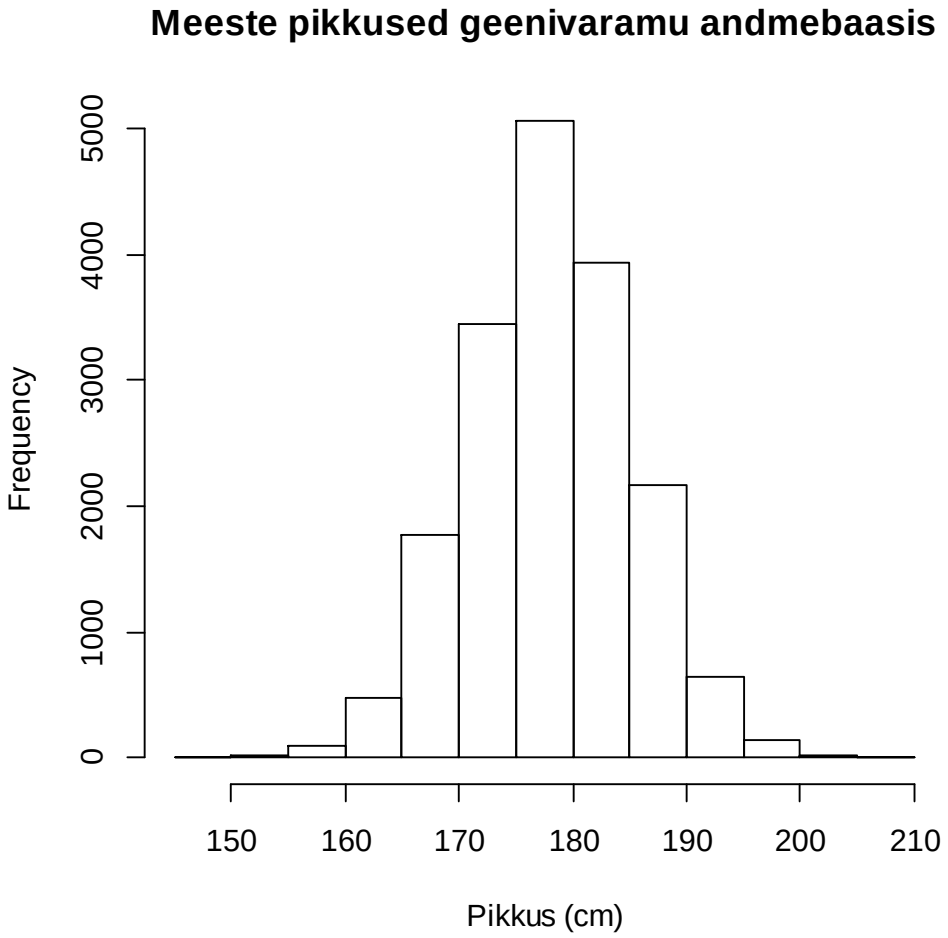
Mis on mudel?

- Annab lihtsustatud pildi tegelikkusest
- Võimaldab uurida mudeli abil kujutatava objekti teatud omadusi

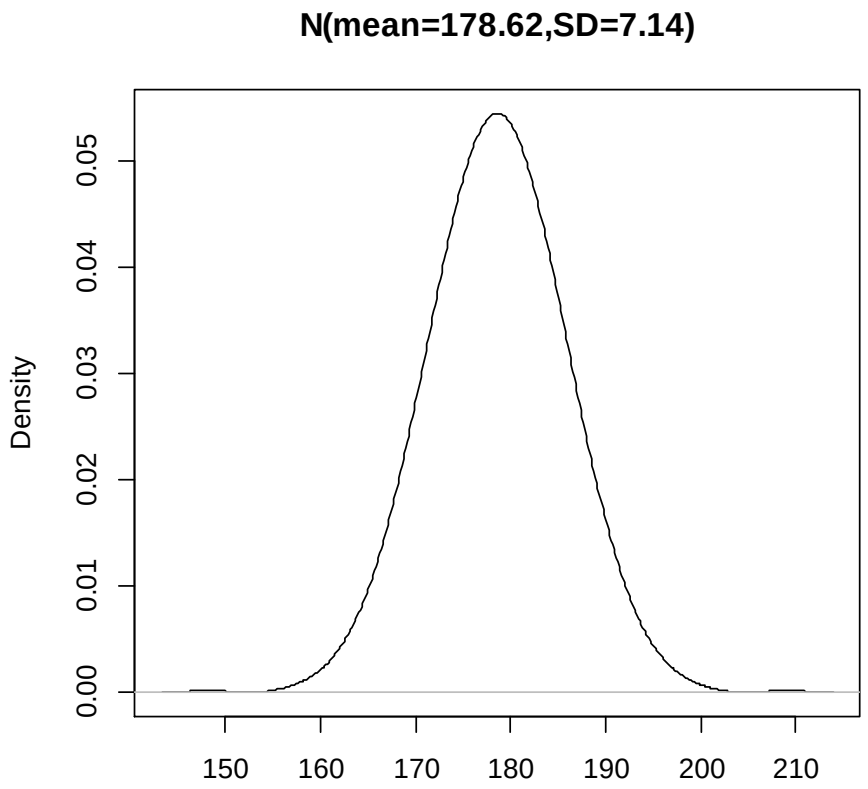
- Statistiline mudel on mudel andmetele
- Statistilise analüüsi eesmärgiks on tihti leida mudel, mis on kooskõlas vaadeldud andmete jaotusega ja/või tunnustevaheliste seostega
- Juba väide tunnuse jaotuse kohta (koos parameetrihinnanguga) kujutab endast mudelit
- Siiski, enamasti kui räägitakse statistilistest mudelitest, mõeldakse nende all seosemudeleid
- Statistiline mudel erineb muudest matemaatilistest mudelitest selle poolest, et ta sisaldab alati juhuslikku komponenti

Näide

Andmed



Mudel



Seosemudelid ja regressioonanalüüs

Statistiline mudel seosele: matemaatiline funktsioon, mis kirjeldab kahe või enama tunnuse vahelist seost – tihti eesmärgiga mõista andmete tekkimise protsessi.

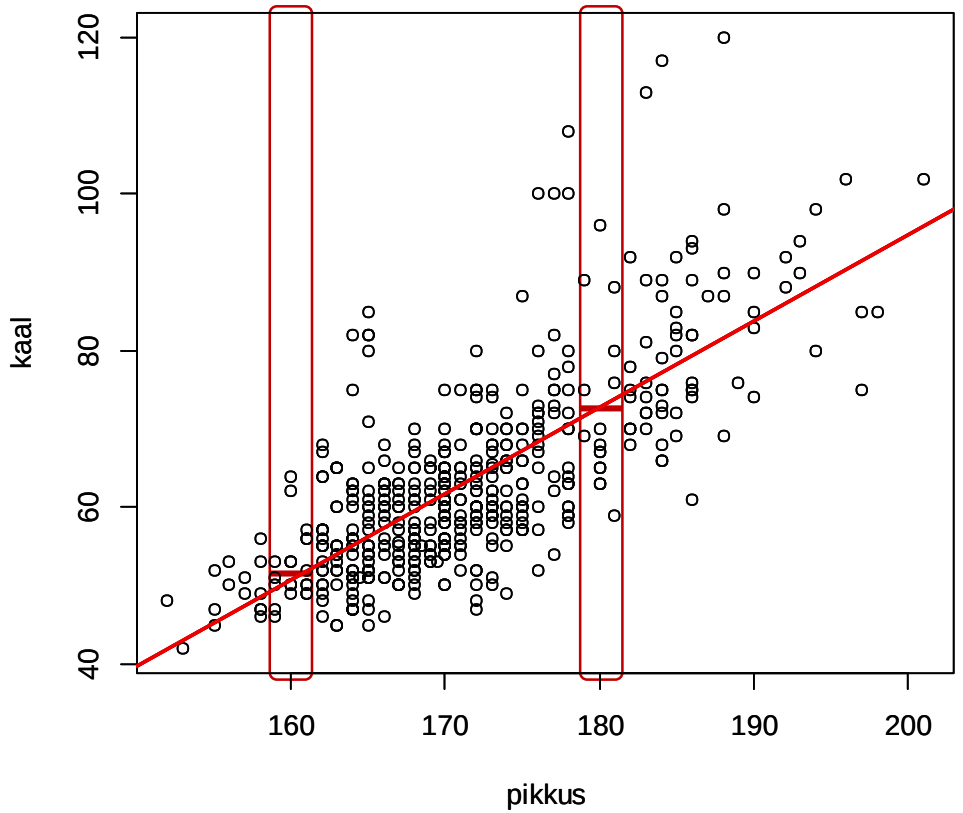
Kõige tavalisemat mudelite hindamise metoodikat nimetatakse regressioonanalüüsiks ja vastavaid mudeleid regressioonimudeliteks.

Regressioonanalüüsis eristatakse tunnuste erinevaid rolle:

- Funktsioontunnus (tähist. Y) – tunnus, mille jaotust soovitakse kirjeldada teis(t)e tunnus(t)e kaudu
- Argumenttunnus (X), ehk ka seletav tunnus (*explanatory variable, covariate*) – tunnus, mille abil soovitakse funktsioontunnuse jaotust kirjeldada
- Argumenttunnuseid võib olla rohkem: X_1, X_2, …, X_k
- Deterministiliku, ehk täpse matemaatilise mudeliga on tegemist siis, kui Y=f(X), st tunnus Y avaldub tunnuse X funktsioonina
- Statistilise mudeli puhul eeldatakse, et $E(Y|X) = f(X)$, st tunnuse Y tinglik keskväärtus (kui X väärtus on teada) avaldub X funktsioonina

Näide: seos pikkuse ja kaalu vahel (arstitudengite andmed)

Keskmine kaal sõltub pikkusest



Sirge võrrand:

$$y = a + bx$$

Aga pikkus ei määra kaalu täpselt! Seega on korrektne kirjutada:

$$\mathrm{E}(Y|X) = a + bX$$

või $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$

kus $\mathrm{E}(\varepsilon_i|X_i) = 0$

ja $i = 1, \ldots, n$

Lihtne lineaarne regressioon

Kahe tunnuse erinevad rollid:

Y – sõltuv tunnus (väljundtunnus, funktsioontunnus, *outcome variable*)

Peamine huvipakkuv tunnus, mille puhul pakub huvi, kui palju on tema varieeruvus seletatav teiste tunnuste abil.

X – sõltumatu tunnus (argumenttunnus, kovariaat, seletav tunnus, *explanatory variable, covariate*).

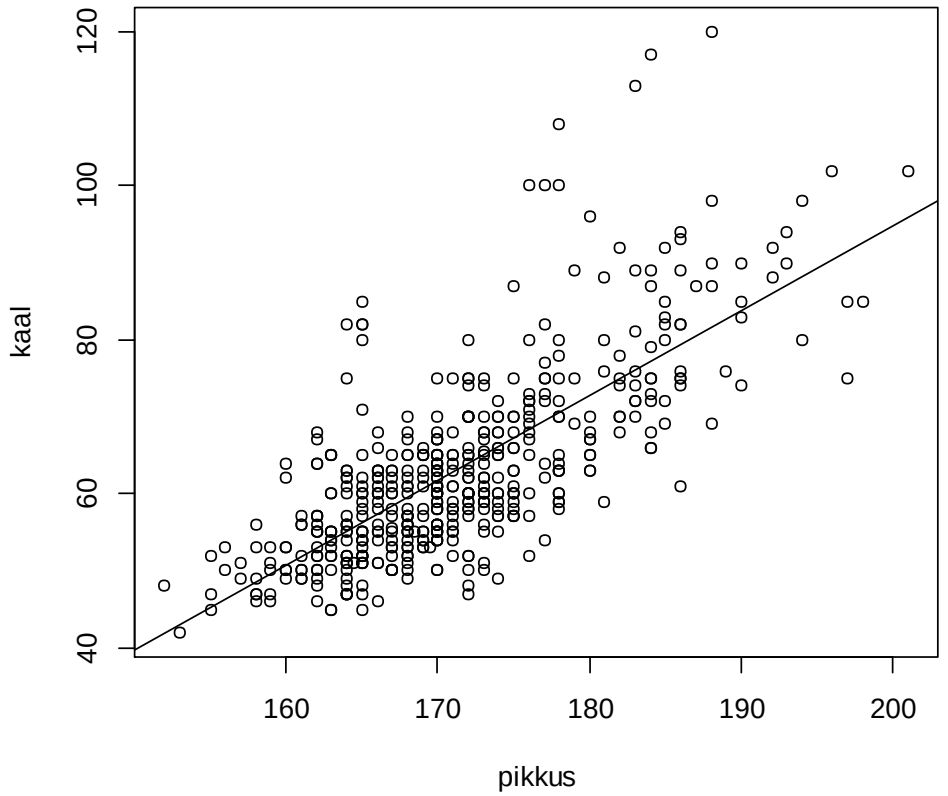
Regressioonimudel: $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

Vabaliige: vajalik nii prognoosimiseks kui selleks, et $\boldsymbol{\varepsilon}$ oleks keskväärtusega 0.

Regressioonikordaja: tema märk ja väärtus määravad X ja Y vahelise seose.
Nullhüpotees: $\boldsymbol{\beta_1 = 0}$

Jääkliige: normaaljaotusega, sõltumatu X-st, keskväärtus 0 ja konstantne varieeruvus

Regressioonimudel ja regressioonisirge



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

β_0, β_1 - parameetrid

$i = 1, \ldots, n$

Analüüsi eesmärk: parameetrite hindamine

regressioonisirge tõus
(tunnuse X kordaja)

Hinnatud mudel:
(kuidas saadi kordajad
-125.5 ja 1.10?)

$$\boldsymbol{Y = -125.5 + 1.10 * X + \varepsilon}$$

Vabaliige (konstant)

Kuidas hinnata mudeli parameetreid?

Vähimruutude meetod: leiame sellised parameetrite väärtused, mis minimiseerivad:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1)^2$$

Kuidas seda tegema peaks?

Kuidas hinnata mudeli parameetreid?

Vähimruutude meetod: leiame sellised parameetrite väärtused, mis minimiseerivad:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - X_i \beta_1)^2$$

Seega vaja lahendada:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \text{ja} \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Vähimruutude meetod

Vaja lahendada:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1)X_i = 0$$

ja

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1) = 0$$

Vähimruutude meetod – maatrikskujul

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1)X_i = 0$$

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta_1) = 0$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \end{pmatrix}$$

$$\mathrm{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix} \qquad \mathrm{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}$$

Võrrandid maatrikskujul:

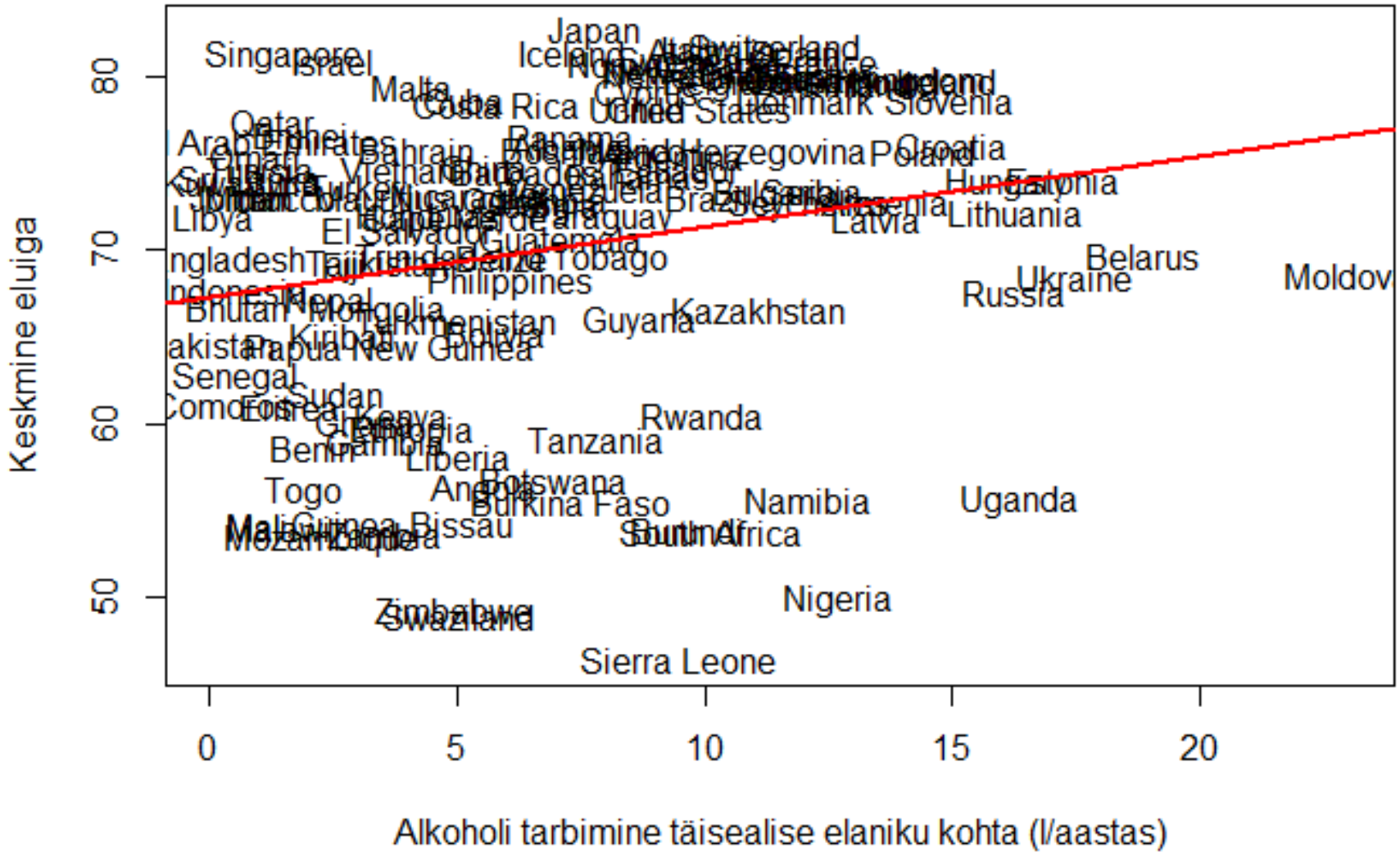
$$\mathrm{X}'(\mathrm{y} - \mathrm{X}b) = 0$$

$$(\mathrm{X}'\mathrm{X})b = \mathrm{X}'\mathrm{y}$$

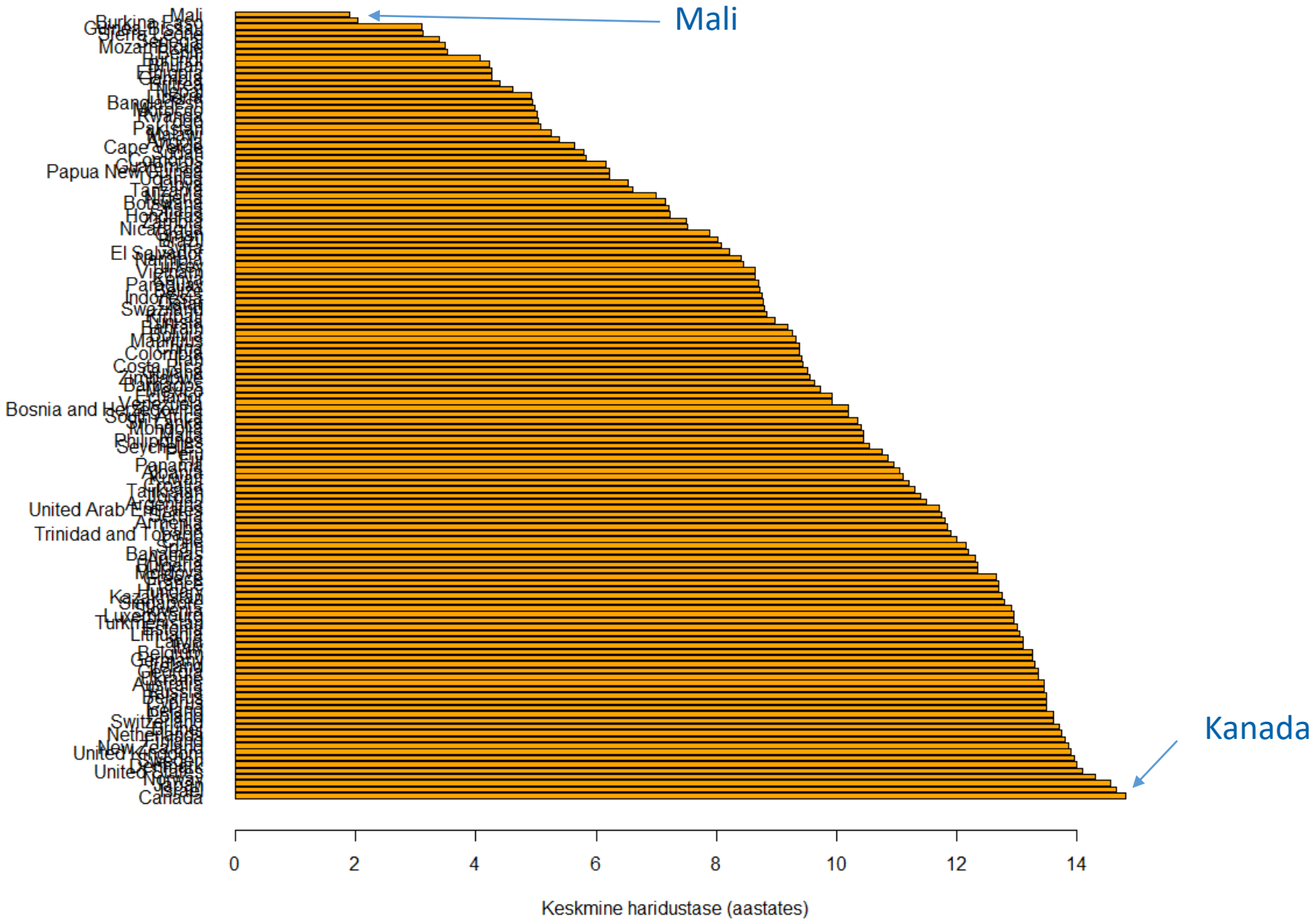
$$b = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

Alkoholi tarbimine ja keskmine eluiga erinevates riikides

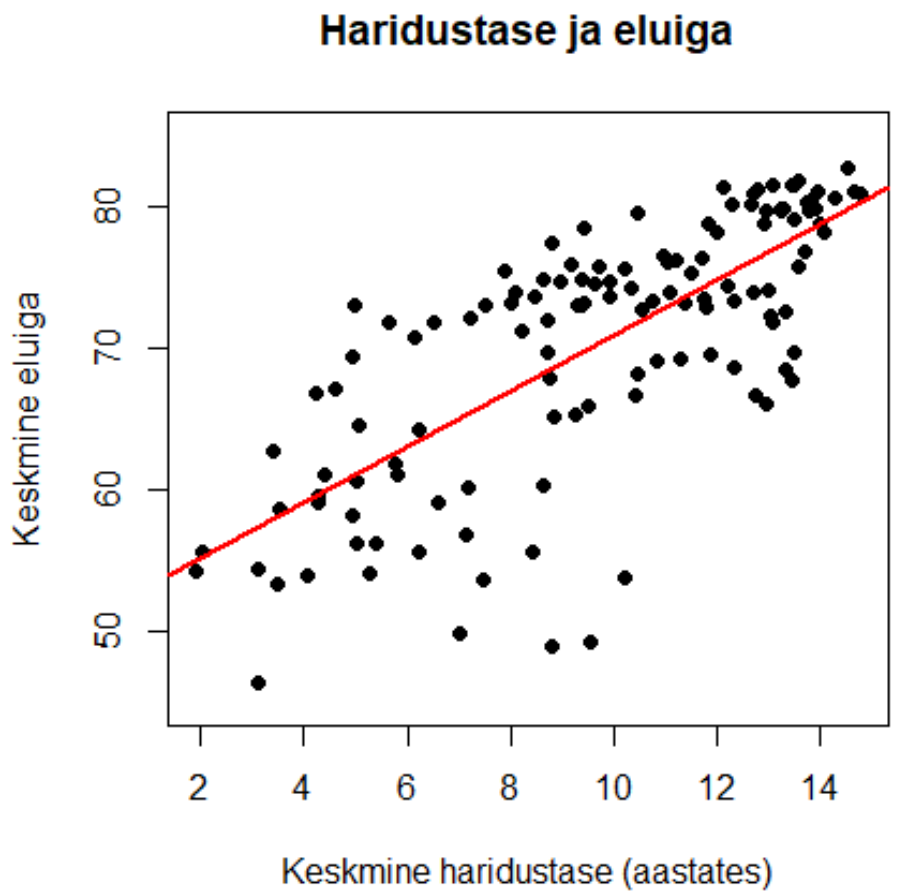
www.gapminder.org – hajuvusgraafik ja regressioonisirge

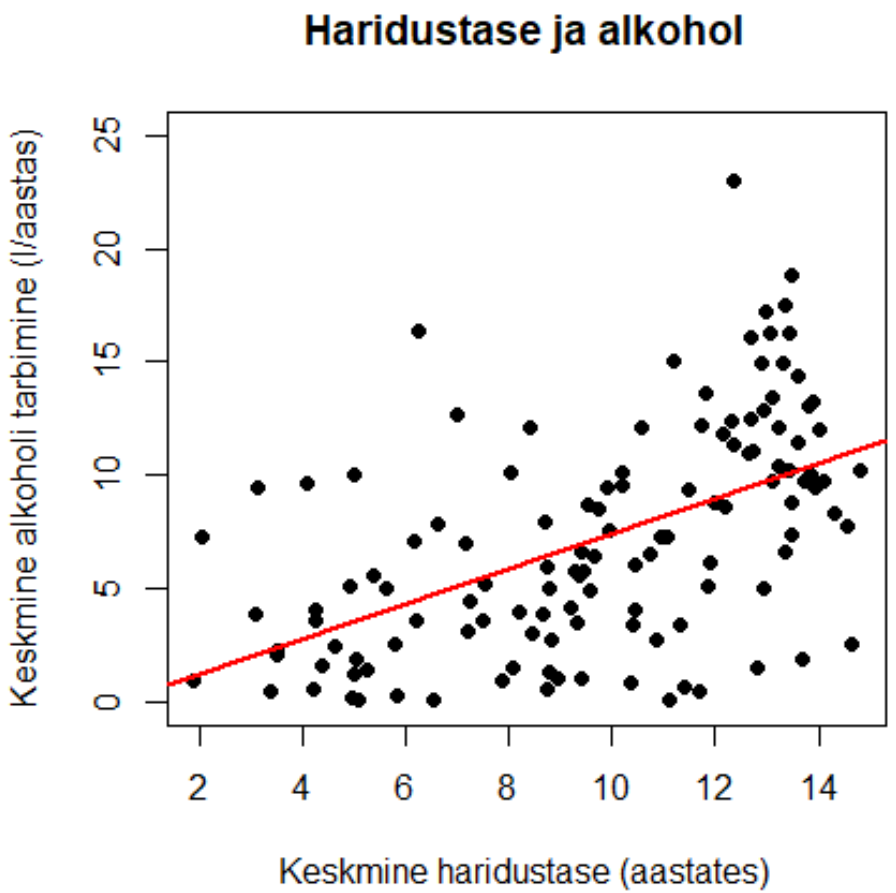


Haridustase riigiti….

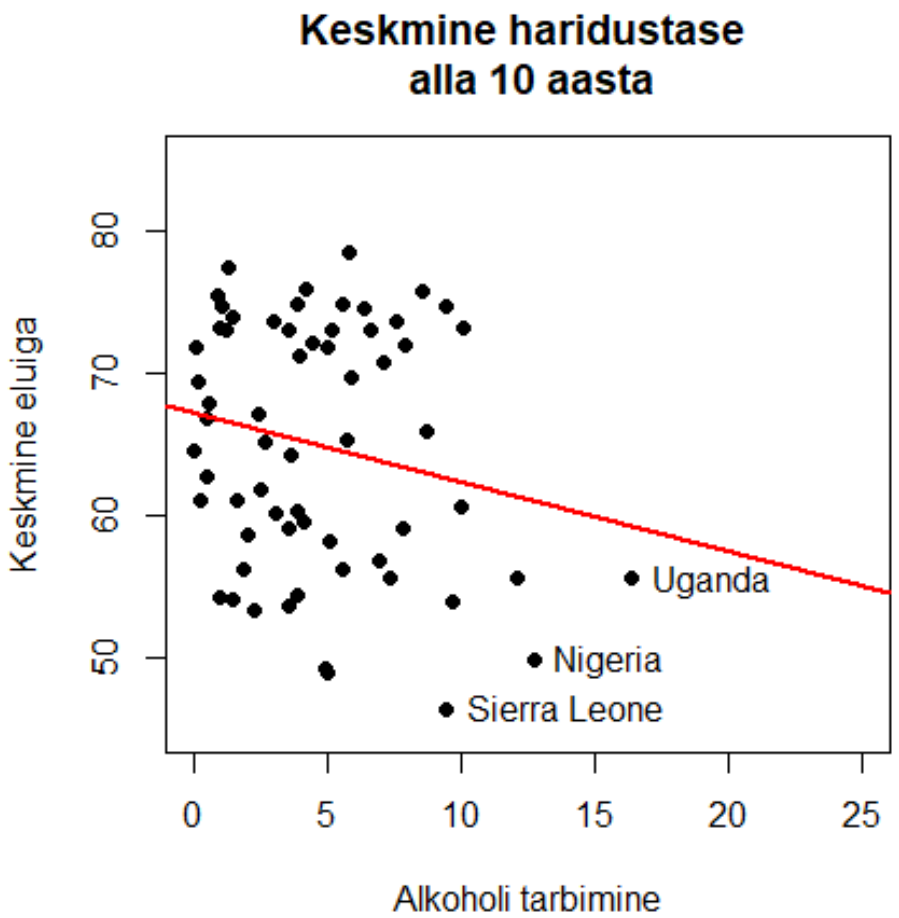


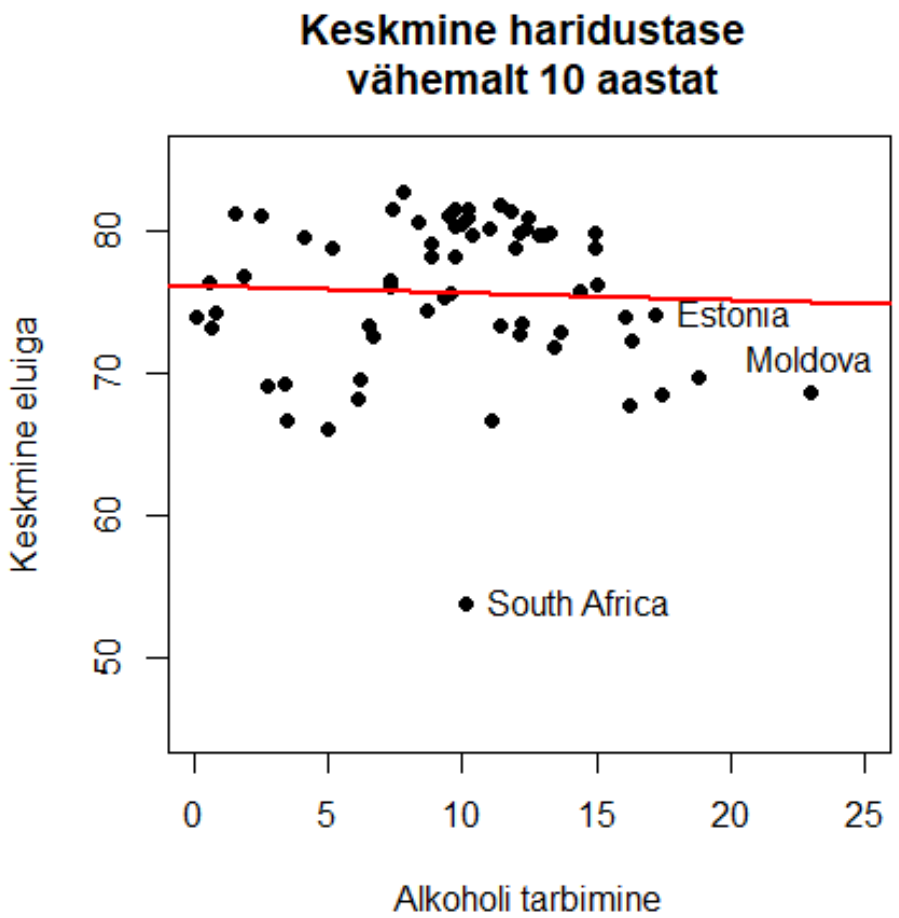
Haridus, eluiga ja alkohol





Alkohol, eluiga ja haridus



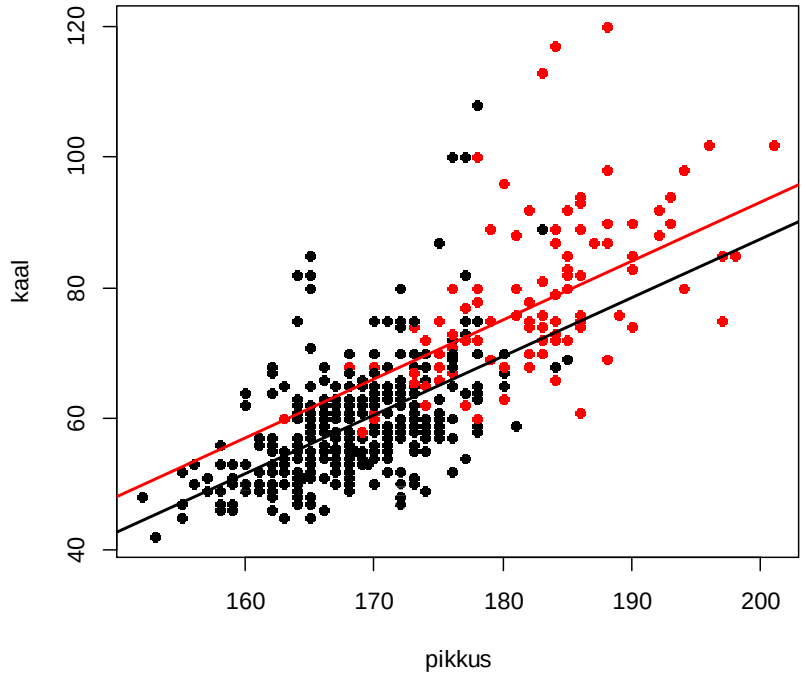


Mitmene regressioon: lisame veel seletavaid tunnuseid

Mitmene lineaarne regressioonimudel:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_k X_k + \varepsilon$$

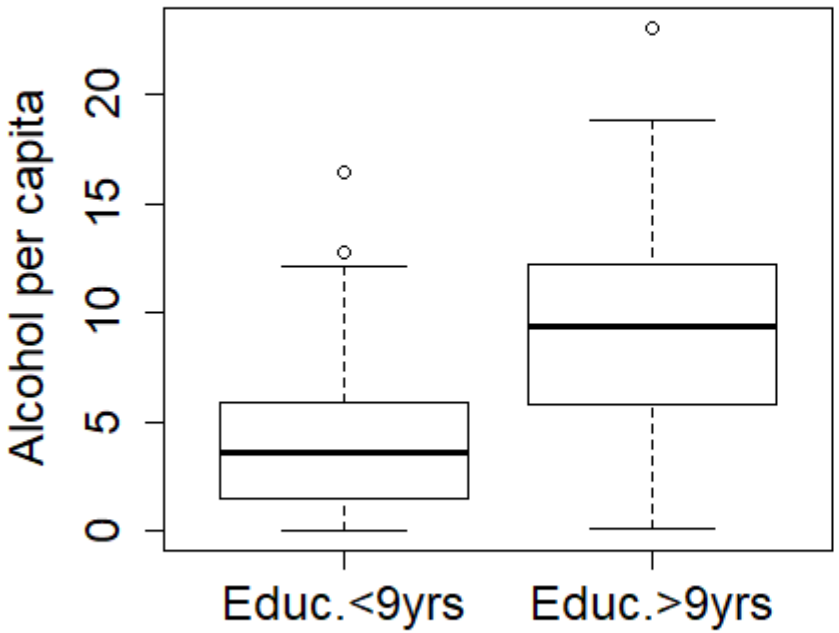
Näide:

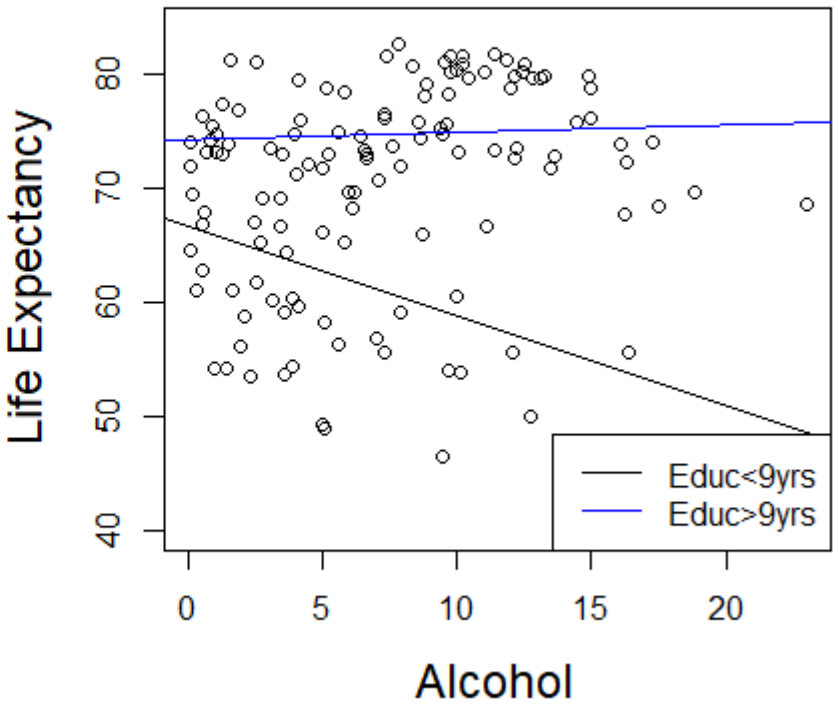


$$\mathbf{kaal} = -98,0 + 0,90 \times \mathbf{pikkus} + 5,53 \times \mathbf{sex} + \ \varepsilon$$

Sugu kodeeritud kui 1- naine, 2-mees

Alkohol, haridus ja eluiga?





Alkoholi seos elueaga sõltub riigi keskmisest haridustasemest – kehvema haridusega riikides paistab mõju olevat tugev ja selgelt negatiivne (alkoholitarbimise ja hariduse koosmõju!)

Mitmene regressioon R abil

```
> summary(lm(kaal~pikkus+sugu+vanus,data=stud))
Call:
lm(formula = kaal ~ pikkus + sugu + vanus, data = stud)

Residuals:
    Min      1Q  Median      3Q     Max
-19.044  -5.012  -1.101   3.610  40.068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -106.45111   10.40384 -10.232  < 2e-16 ***
pikkus         0.90440    0.06347  14.249  < 2e-16 ***
sugu           4.87656    1.26280   3.862 0.000128 ***
vanus          0.42617    0.16612   2.565 0.010606 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.165 on 486 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5618,    Adjusted R-squared:  0.5591
F-statistic: 207.7 on 3 and 486 DF,  p-value: < 2.2e-16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R^2 : nn kirjeldatuse protsent – mudeli prognooside ja Y vahelise korrelatsiooni ruut

Binaarsed argumenttunnused

- Tihti pakuvad huvi tunnused, mis näitavad mingi haiguse vm seisundi olemasolu või puudumist: (väärtus 1 seisundi esinemisel ja 0 selle puudumisel)

- Soovitakse uurida, kuidas haiguse või seisundi esinemise tinglik tõenäosus antud argumenttunnuste väärtuste korral sõltub nendest argumenttunnustest

- Kui uuritavaks tunnuseks on mingi haiguse esinemine, siis neid tinglikke tõenäosusi saab mõnikord tõlgendada kui hinnanguid haiguseriskile ning sellele, millest ja millisel moel see risk sõltub

Logistiline regressioon

- Mudel:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Logaritmitud šansid (log-odds)

Argumendid (Riskifaktorid)

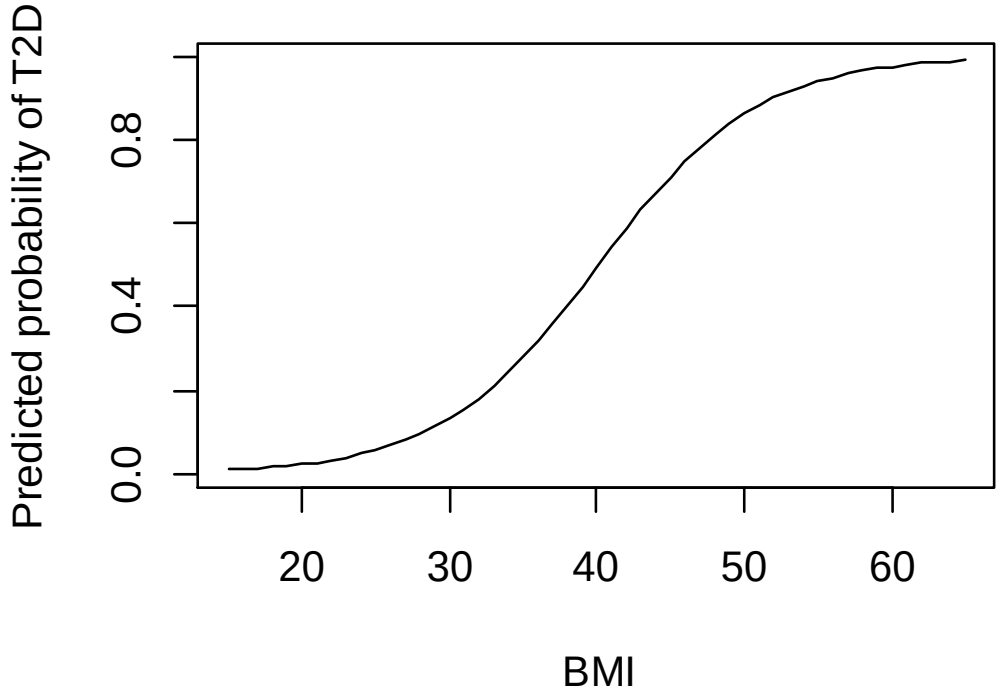
$$p = P(Y = 1 | X_1, X_2, \ldots, X_k)$$

Argumenttunnused võivad olla binaarsed, diskreetsed või pidevad

Parameetrite hindamine: suurima tõepära meetod (eeldatakse, et Y on binoomjaotusega)

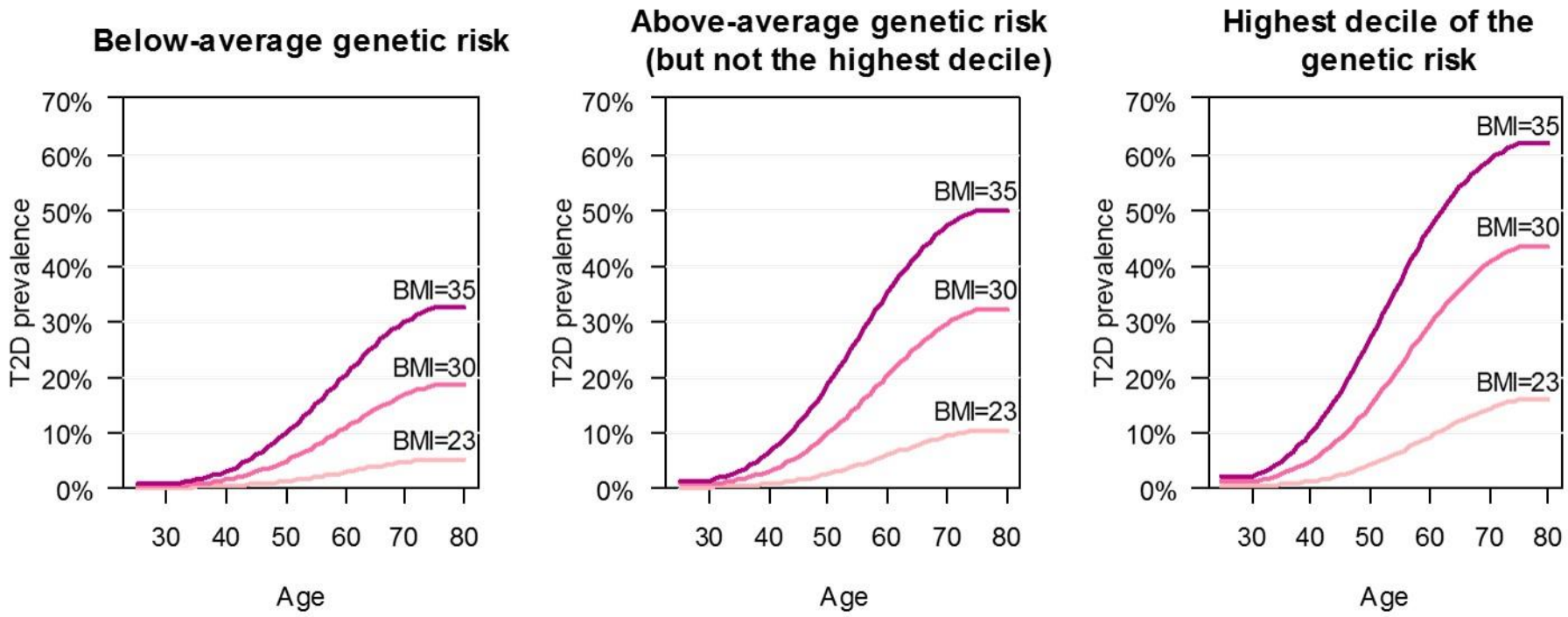
Näide: diabeedi esinemise tõenäosus sõltuvalt kehamassiindeksist

```
mud<-glm(t2d~bmi,family=binomial,data=fen)
pr<-predict(mud,newdata=data.frame(bmi=15:65),type="resp")
plot(15:65,pr,type="l",
xlab="BMI",ylab="Predicted probability of T2D")
```



Logistiline funktsioon garanteerib, et prognoositud tõenäosused jäävad 0 ja 1 vahele

Diabeedi levimus (*prevalence*) sõltuvalt kehamassiindeksist ja geneetilise riski tasemest – logistilise regressioonimudeli prognoosid



Veel mudeleid

- Üldistatud lineaarsed mudelid (nende hulka kuulub ka logistiline regressioon)

$$g(\mathrm{E}(Y)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Sobivad siis, kui Y jaotus kuulub nn eksponentsiaalsesse perre

Erinevad mudelid sõltuvalt funktsioontunnuse või argumenttunnuste tüübist:

- Koosmõjudega mudelid (ühe argumenttunnuse mõju sõltub teise väärtusest)
- Juhuslike mõjudega mudelid (argumenttunnused)
- Mudelid kordusmõõtmistele (funktsioontunnused)
- Mudelid elukestusandmetele (funktsioontunnused)
- Struktuurivõrrandite mudelid (keerukam sõltuvusstruktuur)
- Jne.

Näide: struktuurivõrrandite mudelid