Peatükk 9

Puuduvad andmed

Me võime rahvastikuregistri abil moodustada kõige ilusama juhusliku valimi — aga uuringu lõpuks esineb meil ikka andmestikus puuduvaid väärtuseid. Inimesed keelduvad vastamast, ankeedis on mõni vastus loetamatu või kaugest kolkast pärit proovid osutusid külmutusseadme rikke tõttu kasutuskõlbmatuks. Tegijal juhtub alati — on isegi öeldud, et kui uurimistöö autorid ei kaeble puuduvate väärtuste pärast, siis on neil midagi varjata (näiteks on nad kasutanud sobimatuid meetodeid puuduvatest andmetest vabanemiseks).

Loomulikult tasub alati teha kõik mis võimalik puuduvate väärtuste tekke välistamiseks — näiteks hoolikalt läbi mõeldes, mida patsient, tema raviarst ja meditsiiniasutuse juhtkond antud küsimustikule vastamisest võidavad saab vastamistõenäosust tõsta üle 95%, seevastu uuringutes kust kasu tõuseb eelkõige uuringu läbiviijale võib vastamismäär (ehk vastanute osakaal) olla kõigest 30% või vähem. Ähvardamise või väljapressimise abil saab vastamismäära muidugi tõsta (kui ikka kolmas või neljas kord tullakse jälle küsima, kas ehk nüüd nõustute küsimustele vastama, siis arglikum inimene võibki lõpuks küsimustele vastata — muidu äkki ei saagi sellest tüütust küsitlejast lahti). Aga puuduvaid väärtuseid tekib paratamatult erinevatel põhjustel ja vahel on hea teada, mida nendega peale hakata.

9.1 Puuduvate andmete liigitus

Valimaks sobivat meetodit puuduvate andmetega toimetulekuks kasutatakse mudeleid, mis kirjeldavad andmete olemasolu või puudumist. Me võime näiteks kasutada indikaatortunnust (M), mis näitab seda, kas vaatlus on olemas või mitte, vaata ka tabel 9.1.

Tabel 9.1: Puudumise indikaator.

puudumise indikaator (M)	X_{mis}	X_{obs}	
	haigus	elukoht	sugu
0	1	Tallinn	M
1	-	Tallinn	N
0	0	Tartu	M
0	0	Tartu	N
1	-	Tartu	N
0	0	Tartu	M
0	1	Pärnu	N
⋮	⋮	⋮	⋮

Kui tõenäosus, et vaatlus on puudu, ei sõltu ei olemasolevate tunnuste väärtustest (tunnused, mille väärtused on meil teada: X_{obs}) ega ka selle tunnuse (või nende tunnuste) väärtusest, milles esineb puuduvaid väärtuseid (X_{mis}), siis räägime täiesti juhuslikust puudumisest (*Missing Completely At Random, MCAR*):

$$P(M|X_{mis}, X_{obs}) = P(M).$$

Täiesti juhuslik puudumine tekib näiteks siis, kui osad vaatlused on kaduma läinud seetõttu, et hiired ühe paberankeedi katki närisid. Vaevalt, et hiired otsustasid närimiseks valida ankeeti selle järgi, millised vaatlused või mõõtmistulemused ankeedis on kirjas.

Täiesti juhusliku puudumise korral võime sageli andmete puudumist lihtsalt ignoreerida — võime lihtsalt öelda, et meil on tegemist esialgu kavandatud valimist lihtsalt veidi väiksema valimiga. Aga kõik esindava valimi analüüsiks sobivad töövõtted jäävad kehtima — võime usalduspiire, hinnanguid või statistilisi teste sooritada täpselt nii, nagu me ka muidu võiksime neid teha.

Vahel võib vaatluste puudumise tõenäosus sõltuda vaadeldud või teadolevate tunnuste väärtustest, aga ei sõltu tunnuse enda tegelikust väärtusest. Näiteks kaovad alatihti Pärnu haigla labaratooriumisse saadetud proovid ära (näiteks lohaka kohaliku postiljoni tõttu). Sellisel juhul sõltub puudumise indikaator inimese elukohast (Pärnu elanikul on tõenäolisemalt proovi tulemus teadmata kui tartlasel), aga vaatluse (haiguse olemasolu) puudumine ei sõltu haiguse olemasolust. Sellisel juhul kehtib tingimus

$$P(M|X_{mis}, X_{obs}) = P(M|X_{obs})$$

ja räägitakse juhuslikust puudumisest (*Missing At Random, MAR*).

Juhusliku puudumise korral võib meid huvitava tunnuse (näiteks haiguse olemasolu indikaatori) tinglik jaotus olla erinev olemasolevate vaatluste ja puuduvate väärtuste jaoks (st $P(M|X_{mis}) \neq P(M)$). Pärnus võib haiguse levimus olla suurem kui mujal Eestis. Kui aga Pärnu inimeste proovid kipuvad kaduma minema, siis on olemasolevate vaatluste seas rohkem terveid inimesi kui nende seas, kelle proov läks kaduma ja kelle terviseseisundi kohta meil informatsioon puudub (ja seega oleks antud näite puhul olemasolevate proovide levimus alahinnanguks haiguse tegelikule levimusele Eestis). Seega arvutades olemasolevate vaatluste pealt haiguse levimust — ignoreerides puuduvate väärtuste olemasolu — võib meid viia süstemaatiliselt valede tulemusteni.

Küll aga on vaatluste olemasolu või puudumine tinglikult sõltumatu haigusseisundist peale seda, kui tingimustame elukoha järgi — kui vaatame Pärnust võetud proove, siis nende analüüsitulemuste olemasolu ei sõltu enam haiguse olemasolust inimesel või mitteolemasolust (sest proovid muutuvad kasutuskõlbmatuks pigem lohaka postiljoni tõttu, kes postipaki kiire kohaletoimetamise asemel mõnuleb tunnikese kuumal rannaliival kus kuumuse tõttu varem võetud proov võib labaratoorseks analüüsiks kasutuskõlbmatuks muutuda). Kui aga proovide hukkaminek ei sõltu sellest, kas tegemist on positiivse või negatiivse prooviga — vaid ainult elukohast (kas antud aadressi teenindab lohakas või kohusetundlik postilijon), siis on olemasolevate vaatluste põhjal siiski võimalik leida mõistlikku hinnangut haiguse levimusele näiteks Pärnus (sest Pärnu-piirkonnas, mida teenindab sama postiljon, võiks olemasolevate ja kaotsi läinud proovides haiguse levimus olla sama). Seega on põhimõtteliselt võimalik toodud näite korral leida nihketa hinnangud haiguse levimusele Pärnus, Tartus jne. Saadud nihketa hinannguid mõistlikult kombineerides (näiteks Pärnu, Tartu jne elanike arvu proportsioone kasutades piirkondade levimushinnagute kaalutud keskmist kasutades) on võimalik leida ka adekvaatne hinnang haiguse levimusele Eestis.

Seega juhusliku puudumise korral me ei tohi andmete puudumist niisama ignoreerida (see tooks kaasa vigased ja nihkega hinangud meid huvitavatele suurustele nagu haiguse levimusele), kuid mõistlikult käitudes ja asjakohaseid meetodeid kasutades on siiskli võimalik leida nihketa hinannguid meid huvitavatele suurustele (näiteks levimusele või keskväärtusele).

Kõige tülikamad on need juhud, kui andmete puudumine on mittejuhuslik. Mittejuhuslikust puudumisest (*Not Missing At Random, NMAR*) või informatiivsest puudumisest (*Informative Missingness*) räägitakse siis, kui vaatluse olemasolu või puudumine sõltub puuduvaid väärtuseid sisaldavast tunnusest endast ja olemasolevate tunnuste arvestamine ei hävita seda sõl-

tuvust:

$$P(M|X_{mis}, X_{obs}) \neq P(M|X_{obs}).$$

Sageli võib mittejuhusliku puudumist tingida see, et teatud küsimuse vastust peetakse ühiskonnas taunitavaks või vähemalt küsitletavad ise peavad vastust piinlikuks. Näiteks prostituute küsitledes võime pärida ka selle kohta, kas nad teadaolevalt põevad mõnda seksuaalsel teel levivat haigust. Usutavasti on märksa kergem sellele küsimusele vastata, kui neil pole HIV-nakkust või mõnda muud koledat suguhaigust. Kui aga küsitletaval seksuaalteenuse osutajal on suguhaigus ja ta teab sellest, siis tegelikult ei tohiks ta üldsuse (või vähemalt tema klientide arvatavates) oma tööd edasi teha. Seega öelda, et jah ma tean, et ma põen mõnda sellist haigust on inimesel raske — sestap pigem vastamise asemel loobutakse vastamisest. Tulemuseks on aga vaatluste informatiivne puudumine — kas meid huvitava tunnuse väärtus on puudu sõltub tunnuse X_{mis} väärtusest ja teiste tunnuste järgi tingimustamine (kas tegemist on Tallinna või Narva prostituudiga) seda sõltuvust ei kaota.

Alljärgnevalt vaatame mõningaid strateegiaid, mida saab puuduvaid vaatluseid sisaldava andmestiku analüüsimisel kasutada.

9.2 Sensitiivsusanalüüs

Kui vaatluste puudumine võib olla informatiivne, aga puuduvaid väärtuseid on suhteliselt vähe, siis võib kasutada sensitiivsusanalüüsi abi. Vaatame näiteks järgmist seksuaaltöötajate küsitlemise abil saadud andmestikku. Uuriti, kas küsitletaval on esinenud (või on praegu) suguhaigust. Väärtus 1 tähendab, et on esinenud, 0 tähistab, et küsitletaval pole suguhaiguseid esinenud ja väärtus ? tähistab seda, et küsitletav keeldus vastamast. Hüpoteetilise uuringu tulemused on järgmised: 1; 1; 0; 1; 0; ?; 0; 0; ?; 1. Milline on tõenäosus, et prostituut on põdenud või põeb suguhaigust?

Sensitiivsusanalüüsi korral proovitakse puuduvate väärtuste (?) asemele kirjutada kõiki võimalikke vastuste kombinatsioone. Iga kombinatsiooni korral leitakse hinnang meid huvitavale suurusele (näiteks levimusele). Saadud võimalike hinnangute varieerumispiirkonda kutsutakse ignorantsuspiirkonnaks (*Ignorance region*). Vaata ka tabelit 9.2.

Kuid ignorantsuspiirkond kirjeldab kõigest seda, milline võiks olla hinnang. Meid huvitab aga ju populatsioon, soovime kirjeldada kui täpselt teame uuritava parameetri väärtust populatsioonis (milline on suguhaiguste levimust prostituutide populatsioonis). Ka seda määramatust on võimalik

Tabel 9.2: Kõikmõeldavad puuduvate väärtuste kombinatsioonid (mustrid) annavad ignorantsuspiirkonnaks 40%-60%.

Muster	Valim										
1	1	1	0	1	0	**0**	0	0	**0**	1	$\widehat{Pr} = 40\%$
2	1	1	0	1	0	**1**	0	0	**0**	1	$\widehat{Pr} = 50\%$
3	1	1	0	1	0	**0**	0	0	**1**	1	$\widehat{Pr} = 50\%$
4	1	1	0	1	0	**1**	0	0	**1**	1	$\widehat{Pr} = 60\%$

kirjeldada. Iga puuduvate väärtuste mustri korral võime leida ju usaldusintervalli meid huvitavale parameetrile. Nende usaldusintervallide ühend annab tulemuseks määramatuspiirkonna (*Uncertainty region*). Näiteks erinevatele puuduvate väärtuste võimalikele mustritele vastavad 95%-usaldusintervallid ja tulemuseks saadud määramatuspiirkond on esitatud tabelis 9.3.

Tabel 9.3: Määramatuspiirkonna leidmine, mustrid nagu toodud tabelis 9.2.

Muster	95%-usaldusintervall
1	[0,122 . . . 0,738]
2	[0,187 . . . 0,813]
3	[0,187 . . . 0,813]
4	[0,262 . . . 0,878]
95%-määramatuspiirkond:	[0,122 . . . 0,878]

Kui puuduvaid väärtuseid sisaldav tunnus on pidev tunnus (näiteks kaal: osad — arvatavasti paksemad — tüdrukud keelduvad ütlemast oma kaalu) siis pole muidugi võimalik asendada puuduvaid väärtuseid kõikvõimalike kaalu väärtustega. Küll aga on võimalik teha näiteks sellist analüüsi, kus sõltuvalt mingist hästi arusaadavast parameetrist (näiteks: mittevastavate tüdrukute keskmine kaal) näidatakse meid huvitava tulemuse muutumist (usaldusintervall tüdrukute kaalu keskväärtusele või tüdrukute ja poiste kaalude erinevusele).

Paar näidet määramatuspiirkonna kasutamise kohta võid leida näiteks Vansteelandt *et al.*, 2006; Vansteelandt ja Goetghebeur, 2001.

9.3 Mitmene imputeerimine

Kui vaatluste mittejuhusliku puudumise korral on sensitiivsusanalüüs üks väheseid võimalusi midagigi oma andmetest kätte saada, siis andmete juhusliku puudumise korral on juba mitmeid võimalusi, kuidas korrektset analüüsitulemust saavutada. Üks kõige üldisem ja väga paljude erinevate analüüside korral kasutatav meetod on puuduvate väärtuste mitmene imputeerimine.

Alljärgnevalt kirjeldame mitmese imputeerimise meetodi ideed.

Kõigepealt leitakse puutuvaid väärtuseid sisaldava tunnuse või tunnuste (X_{mis}) tinglik jaotus, tingimusel et on teada teadaolevate tunnuste (X_{obs}) väärtused: $F_{X_{mis}|X_{obs}}$. Kasutades saadud tinglikku jaotust genereeritakse puuduvate väärtuste asemele juhuslikud suurused sellest tinglikust jaotusest (kui haiguse indikaator on puudu ühel pärnakal siis genereeritakse haigustunnuse väärtus binoomjaotusest lähtuvalt haiguse levimusest Pärnus). Seega kui i. inimesel on tunnuse X_{mis} väärtus puudu, siis genereerime puuduva väärtuse jaotusest: $X_{mis;i} \sim F_{X_{mis}|X_{obs}=X_{obs;i}}$.

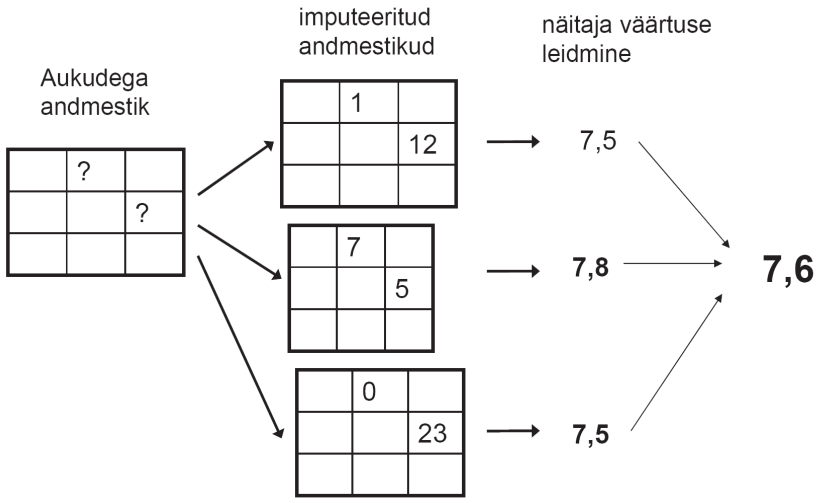
Sellisel viisil saame tekitada andmestiku, kus puuduvate väärtuste asemel on mingid genereeritud juhuslikud suurused. Sellist andmestikku on lihtne analüüsida mistahes meetodi abil, mida soovime kasutada (regressioonanalüüs, logitiline regressioon, ...).

Järgnevalt kordame kogu protseduuri mitmeid kordi — genereerime puuduvate väärtuste asemele uued juhuslikud suurused ja teostame oma analüüsi uuesti. Iga kord salvestame meid huvitava hinnangu (näiteks keskmise või regressioonmudeli mingi kordaja väärtuse). Kuna auke samas andmestikus uuesti juhuslike suurustega täites saame veidi teistsuguse andmestiku siis tuleb ka meid huvitava statistiku väärtus veidikene teistsugune kui varem. Sellisel viisil täidame andmestikus olevaid auke (puuduvaid väärtuseid) mitmeid kordi (näiteks genereerime 10 andmestikku ja teostame kümme analüüsi). Saadud hinnangud keskmistatakse saamaks lõplikku hinnangut meid huvitavale suurusele, vaata ka joonist 9.1.

Saadud hinnangu täpsust tuleb ka kirjeldada. Siin tuleb appi valem, mis kirjeldab kuidas tingliku dispersioone ja keskväärtuseid saab kasutada dispersiooni leidmiseks. Nimelt saame tavapäraste meetoditega hinnata hinnangu dispersiooni, kui fikseerime mingi imputeeritud andmestiku (IMP), ehk lihtne on mingit konkreetset puuduvate väärtusteta andmestikku kasutades leida hinnang $\mathrm{E}(\hat{\beta} \mid IMP)$ ja tema dispersioon $\mathrm{D}(\hat{\beta} \mid IMP)$. Seejärel saame aga leida hinnangu dispersiooni esialgse, puuduvaid väärtuseid sisaldava andmestiku jaoks:

$$\mathrm{D}(\hat{\beta}) = \mathrm{ED}(\hat{\beta} \mid IMP) + \mathrm{DE}(\hat{\beta} \mid IMP),$$

Joonis 9.1: Mitmene imputeerimine



kus ED($\hat{\beta} \mid IMP$) on leitud dispersioonihinnangute keskmine ja DE($\hat{\beta} \mid IMP$) on leitud hinnangute dispersioon. Praktikas tavaliselt kohendatakse saadud arvutusvalemit veidi arvestamaks võimaliku veaga, mida teeme kui imputeerimiseks vajalike juhuslike suuruste genereerimisel kasutame jaotuse $F_{X_{mis}|X_{obs}}$ asemel hinnatud jaotust $\hat{F}_{X_{mis}|X_{obs}}$.

Kuidas imputeerimist reaalselt näiteks R-is saab kasutada? Üheks parimaks töövahendiks puuduvate andmetega ringikäimisel on lisamoodul mice.

Vaata ja proovi läbi kaks alltoodud mice-i kasutusnäidet ja proovi siis iseseisvalt lahendada peatüki lõppu lisatud ülesannet. Põhjalikuma näite võid leida Katitas, 2019.

Näide 1

```
# Algandmed − puuduvate väärtustega

andmedpisi=data.frame(
  kaal=c(74.0, 100.0, NA, NA, 57.0, 54.0, NA,  92.0,
                                          47.0, 82.0, NA ),
  sugu=factor(c("M", "N", "N", "N", "N",
    "N", "N", "M", "N", "M", "N")),
```

```
  lemmikvarv=factor(c("roheline", "punane", NA, "roosa",
    "punane","roheline", "punane", "roosa", NA,
    "roheline", "punane")))

# Tekitame viis imputeeritud väärtustega andmestikku
imp=mice(andmedpisi, m=5)

# vaatame 2. ja 3. imputeeritud andmestikku
complete(imp, 2)
complete(imp, 3)

# Hindame meid huvitavad suurused kõigis viies andmestikus
# Antud juhul huvitavad meid näiteks lineaarse mudeli parameetrid.
hinnangud <-with(data=imp, exp=lm(kaal~lemmikvarv))

# Koondame saadud 5 komplekti hinnanguid üheks korrigeeritud hinnanguks.
summary(pool(hinnangud), conf.int=TRUE)

# ----------------------------------------
# Variant 2: soovi korral või vajadusel võib
# täpsustada, milliste tunnustega on meil tegemist ehk
# täpsemalt: millist mudelit kasutatakse tingliku jaotuse jaoks.
# Järgnevalt ütleme, et:
#         1. tunnus on normaaljaotusega,
#         2. tunnust pole vaja imputeerida
#         3. tunnus on diskreetne (töötab ka mudue
imp=mice(andmedpisi, m=5, method=c("norm", "", "polyreg"))

# Vaata: siin on näha, millise tunnuse imputeerimiseks
# millist mudelit kasutati.
summary(imp)

# Vaata, kuidas on muutunud imputeeritud kaalud!
complete(imp, 2)

# Ülejäänud jääb samaks: hindame mudelid ja kogume
# saadud hinnangud üheks hinnanguks kokku:
hinnangud <-with(data=imp, exp=lm(kaal~factor(lemmikvarv)))
summary(pool(hinnangud), conf.int=TRUE)
```

Näide 2

Järgnevalt tekitame ühe simuleeritud andmestiku. Selles genereeritud andmestikus on Tartus haiguse levimus 0,5 ja Tallinnas 0,2. Teisi Eesti piirkondi ei vaadelda - uuritavaks populatsiooniks ongi vaid Tartu ja Tallinna elanikud. Kirjeldatud populatsioonist moodustavad 18% tartlased ja 82% on tallinnlased.

```
pr_tartu=0.5
pr_tallinn=0.2
p_tartu=0.18

# Juhusliku valimi suurus.
n=10000

set.seed(1)

# Mitu tartlast sattub valimisse?
n_tartu=rbinom(1, p_tartu, size=n)

# Küsitletavate elukohad:
piirkond=rep(c("Tallinn", "Tartu"), c(n-n_tartu, n_tartu))

# Kas valimisse sattunud inimesel on haigus (1) või pole haigust (0):
haige=rbinom(n, size=1, prob=pr_tartu*(piirkond=="Tartu")+
                            pr_tallinn*(piirkond=="Tallinn"))

# Tekitame mittevastamise. Tartlased on paremad vastajad:
mittevastaja=rbinom(n, size=1, prob=0.05*(piirkond=="Tartu")+
                                    0.7*(piirkond=="Tallinn"))
haige[mittevastaja==1]=NA

# Genereeritud andmestik:
andmed=data.frame(h=haige, koht=factor(piirkond))
```

Vaata kuidas andmeid genereeritakse. Milline on haiguse tegelik levimus uuritavas populatsioonis (millesse kuuluvad kõik tartlased ja tallinnlased)?

```
# Soovime hinnata haiguse levimust meid huvitavas populatsioonis.
# Esmalt valed analüüsid:
mean(haige, na.rm=T)
```

```
# Usaldusintervall levimusele:
# t-test:
t.test(haige)

# sama lineaarse mudeli abil:
mudel_lm=lm(haige~1)
summary(mudel_lm)
confint(mudel_lm)

# Või kasutades logistilist regressiooni :
m1=glm(haige~1, family=binomial())
summary(m1)
exp(coef(m1))/(1+exp(coef(m1)))
# või, alternatiivselt :
binomial()$linkinv(coef(m1))
# Usaldusintervall:
binomial()$linkinv(confint(m1))
```

Kas tegelik haiguse levimus jäi kasvõi ühtegi neist leitud usaldusintervallidest?

Proovime teostada korrektsema analüüsi kasutades mitmest imputeerimist:

```
library(mice)

# Imputeeri 1. tunnust kasutades logistilist regressiooni
#  2. tunnust ära imputeeri (seal pole puuduvaid väärtuseid)
imp <- mice(andmed, m=10, method=c("logreg", "") )


# Usaldusintervall, versioon 1:
fit1 <-with(data=imp, exp=lm(h~1))
summary(pool(fit1), conf.int=TRUE)

# Usaldusintervall, versioon 2 (logistiline regression ):
fit2 <-with(data=imp, exp=glm(h~1, family=binomial()))
summary(pool(fit2), conf.int=TRUE)
binomial()$linkinv(unlist(summary(pool(fit2), conf.int=TRUE)[7:8]))
```

Kas korrektsemad usalduspiirid sisaldasid tegelikku levimust?

Ülesanne

Loe sisse tartu ülikooli tudengite küsitlemisel saadud tudengite andmestik:

```
load(url("http://www.ms.ut.ee/mart/andmeteadus/tudengid.RData"))
```

Sisseloetud andmestikus (tudengid) on tunnused nagu
kaal - tudengi kaal (kg)
sugu - tudengi sugu (1 - naine; 2 - mees).

Soovime hinnata meest- ja naistudengite kaalude keskväärtuste erinevust.

Kui puuduvaid andmeid poleks või kui ignoreriksime puuduvaid väärtuseid võiksime kaalude keskväärtuseid võrrelda näiteks nii:

```
m=lm(kaal~sugu, data=tudengid)
summary(m)
confint(m)
```

ja näeksime, et hinnanguliselt on meeste kaalude keskväärtus 17,308 kg suurem naiste kaalude keskväärtusest (95%- usaldusintervall keskväärtuste erinevusele: 15,570...19,046).

Kui aga eeldada, et puuduvate andmete tekkemehhanism on juhuslik (kui tegemist oleks juhusliku puudumisega), siis võiksime kasutada mitmest imputeerimist.

Sinu ülesanne: Kasutades mitmest imputeerimist leia hinnang meeste ja naiste kaalude keskväärtuste erinevusele. Kas see tuli suurem või väiksem varem leitud erinevuse hinnangust? Kas hinnang keskväärtuste erinevusele muutus sinu arvates õiges suunas? Miks sa nii arvad? Selgita, kas praegu võiks olla tegemist juhusliku puudumisega? Põhjenda oma otsust: Miks sa nii arvad?

bibliograafia

[1] Aycan Katitas. *Getting Started with Multiple Imputation in R*. 2019. URL: `https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/` (vaadatud 19.04.2022).

[2] Stijn Vansteelandt ja Els Goetghebeur. "Analyzing the sensitivity of generalized linear models to incomplete outcomes via the IDE algorithm". *Journal of computational and graphical statistics* 10.4 (2001), lk. 656–672.

[3] Stijn Vansteelandt *et al.* "Ignorance and uncertainty regions as inferential tools in a sensitivity analysis". *Statistica Sinica* (2006), lk. 953–979.