Biostatistika

Puuduvad andmed

Tartu tähtsusest

Kas noortelinn Tartu tähtsus Eesti noorte jaoks kasvab või kahaneb? Vaatame, milline on Eestis elava noore inimese (laps vanuses 0-4 aastat) tõenäosus elada Tartus nüüd ja vanasti:

Aasta	Lapsi Eestis	lapsi Tartus	lapse elukoht (linn/maakond) teadmata
2023	53 433	4 780	0
2018	50 200	5 158	13

Kuidas muutub ajas lapse tõenäosus elada Tartus? Miks meil (täpsemalt: statistikaametil) esineb andmestikus puuduvaid väärtuseid? Milline võiks olla puuduvate andmete tekkemehhanism (MCAR/MAR/NMAR)?

Iseloomusta tõenäosuste muutust kasutades riskide vahet (RD)!

Kas antud juhul tuleks või ei tuleks leida ka usaldusintervall riskide vahele?

Milline tuleks antud usaldusintervall, kui arvestame puuduvate vaatluste (esineb lapsi, kelle elukoht Eestis pole teada) olemasoluga?

> **Märkus:** usaldusintervalli riskide vahele saab suurte valimite korral leida prop.test käsuga, väikeste valimite korral – alla 100 vaatluse grupi kohta – tasuks kasutada täpset meetodit, näiteks käsku BinomCI lisamoodulist ExactCIdiff – mis on väga aeglane suuremate valimite korral. Näiteks kui ravime 30 inimest kasutades ravi A ja 30 inimest kasutades ravi B, ja ravi A saanutest sureb 2 ning ravi B saanutest sureb 10, siis usaldusintervallid riskide vahele on leitavad järgmiselt:

```
# Ligikaudsed:
prop.test(x=c(2, 10), n=c(30, 30))
# -0.49084738 -0.04248596

# Täpsed (arvutab paar minutit!):
library(ExactCIdiff)
BinomCI(n1=30, n2=30, x=2, y=10, precision=0.001)
# -0.483 -0.065
```

Mitmene imputatsioon:

Loeme sisse näidisandmestikud (puuduvaid väärtuseid sisaldav andmestik andmed ja võrdluseks andmestik, kus kõik vaatlused on alles, andmed_koik):

```
load(url("https://www-1.ms.ut.ee/mart/biostat/MI.RData"))
```

Vaata ka anmdeid:

```
andmed[1:11,]
```

Märka, et andmestikus esineb ka puuduvaid väärtuseid. Kas puudumine võiks olla MAR või MCAR või… ? Vaata ka järgmiste käskude tulemusi:

```
prop.table(table(andmed$maakond, is.na(andmed$y)),1)
chisq.test(table(andmed$maakond, is.na(andmed$y)))
```

Soovime uurida, kuidas nooruki vanus mõjutab y-tunnust. Võrdle (tavaelus saad teostada neist analüüsidest vaid ühte…):

```
m=lm(y~vanus, data=andmed)
summary(m)
confint(m)

m2=lm(y~vanus, data=andmed_koik)
summary(m2)
confint(m2)
```

Milline on seos vanuse ja y-tunnuse vahel? Kas oleksid jõudnud õige järelduseni andmestikku andmed kasutades?

Imputeerime puuduvad väärtused enne analüüsi:

```
library(mice)
# Mice 'i kasutamise eelduseks on see, et faktortunnused on andmestikus
# eelnevalt tehtud faktortunnusteks (mice peab mõistma, mis tüüpi tunnusega
# on tegemist):
andmed$maakond2=factor(andmed$maakond)
andmed$vanus2=factor(andmed$vanus)
andmed2=andmed[,3:5]


# Tekitame 10 imputeeritud andmestikku.
imp=mice(andmed2, m=10, method=c("norm", "", "logreg"))
# Võrdleme saadud andmestikke esialgsete andmetega:
andmed2[1:10,]
complete(imp,1)[1:10,]
complete(imp,2)[1:10,]

# Teostame soovitud analüüsid kasutades kõikki tekitatud andmestikke:
analyysid=with(imp, lm(y~vanus2))
pool(analyysid)
# Ja võtame saadud 10 tulemust tagasi kokku üheks tulemuseks;
summary(pool(analyysid))
summary(pool(analyysid), conf.int=TRUE)
```