Laskennallisten menetelmien ja bayesilaisuuden perusteet

Syksy 2020

Harjoitustyö

Maria Seppänen, maria.seppanen@tuni.fi

Pieni analyysi Tampereen vuonna 2016 myydyistä kaksioista

Käytetty data

Käytin analyysissä dataa Tampereen vuonna 2016 myydyistä kaksioista. Data on peräisin Moodlen Tutkimusmenetelmien työkalupakista linkin ”Tilastotieteen peruskurssien harjoitusaineistoja 2003–2018” alta. Analyysiin on käytetty aineistosta kohtia ”Rakennusvuosi”, ”Neliöhinta”, ”Parveke”, ”Neliöt” ja ”Hinta”. Koko datassa on 564 havaintoyksikköä.

Käytetyt menetelmät

Datan analysoimiseen olen käyttänyt muutamaa eri menetelmää ja R-ohjelmistoa. Tuloksien ohesta tulee löytymään R-koodi, jota tulosten saamiseen on käytetty, sekä muutamia kuvaajia selkeyttämään visuaalisesti. Ohessa pienet esittelyt kustakin käytetystä menetelmästä.

Bootstrap

Bootstrap-menetelmässä alkuperäisestä otoksesta samoin jakautuneita riippumattomia satunnaismuuttujia generoidaan samankaltaisia satunnaisotoksia. Näistä bootstrap-otoksista voidaan sitten laskea estimaatit, ja estimaateilla estimoida otoksen jakaumaa, tai erinäisiä tunnuslukuja, kuten tämän analyysin tapauksessa otoskeskiarvoa.

Permutaatiotesti

Tarkoituksena on selvittää, onko otoksien keskiarvojen erotus tilastollisesti merkitsevä. Tämä tehdään luomalla suuri määrä satunnaisia permutaatioita alkuperäisistä otoksista, ja laskemalla näiden otosten keskiarvojen erotukset. Näiden avulla voidaan arvioida, onko alkuperäisten otoskeskiarvojen erotus tilastollisesti merkitsevä vai ei esimerkiksi p-testillä.

Monte Carlo -testi

Tällä menetelmällä voi tutkia, onko jokin data satunnaisesti levittynyt vai onko datassa näkyvillä jotain viitteitä ei-satunnaisuudesta. Testissä simuloidaan täysin satunnaisesti jakautuneita alkuperäisen otoksen kokoisia ja samalla alueella olevia satunnaisotoksia, ja verrataan pisteiden etäisyyksiä. Tarkastelemalla satunnaisotosten pisteiden etäisyyksien jakaumaa, voidaan päätellä, onko alkuperäisen otoksen pisteiden etäisyys linjassa satunnaisesti jakautuneiden pisteiden jakauman kanssa.

Datan analysointi ja sen tulokset

Tutkin ensiksi Bootstrap-menetelmän avulla neliöiden, hinnan ja neliöhinnan keskiarvojen bootstrap-estimaatteja. Otoksen estimaattina käytin 25 % trimmattua keskiarvoa alkuperäisistä otoksista. Laskin myös harhan ja sain lopulta seuraavanlaiset harhattomat keskiarvon bootstrap-estimaatit:

Neliöt: 52,7 m^2,

Hinta: 132 854 euroa,

Neliöhinta: 2865,16 euroa/m^2.

```
> ## BOOTSTRAP ##
>
> bootstrap.estimate <- function(x) {
+    b<-numeric(10)
+    for (i in 1:10){b[i]<-mean(sample(x,8,replace=TRUE),
+    trim=0.25)}
+    mean.est <- sum(b)/10
+    return(mean.est)
+ }
> bmean.Neliöt <- bootstrap.estimate(Neliöt)
> # Bias-corrected estimaatti
> mean(Neliöt)-(mean(Neliöt, trim=0.25)-bmean.Neliöt)
[1] 52.86101
>
> bmean.Hinta <- bootstrap.estimate(Hinta)
> # Bias-corrected estimaatti
> mean(Hinta)-(mean(Hinta, trim=0.25)-bmean.Hinta)
[1] 132854.7
>
> bmean.Neliöhinta <- bootstrap.estimate(Neliöhinta)
> # Bias-corrected estimaatti
> mean(Neliöhinta) - (mean(Neliöhinta, trim=0.25)-bmean.Neliöhinta)
[1] 2865.163
```

Seuraavaksi tutkin, vaikuttaako parveke asunnon neliöhintaan tilastollisesti merkitsevästi 5 prosentin merkitsevyystasolla. Käytin tähän permutaatiotestiä, jonka hypoteeseina

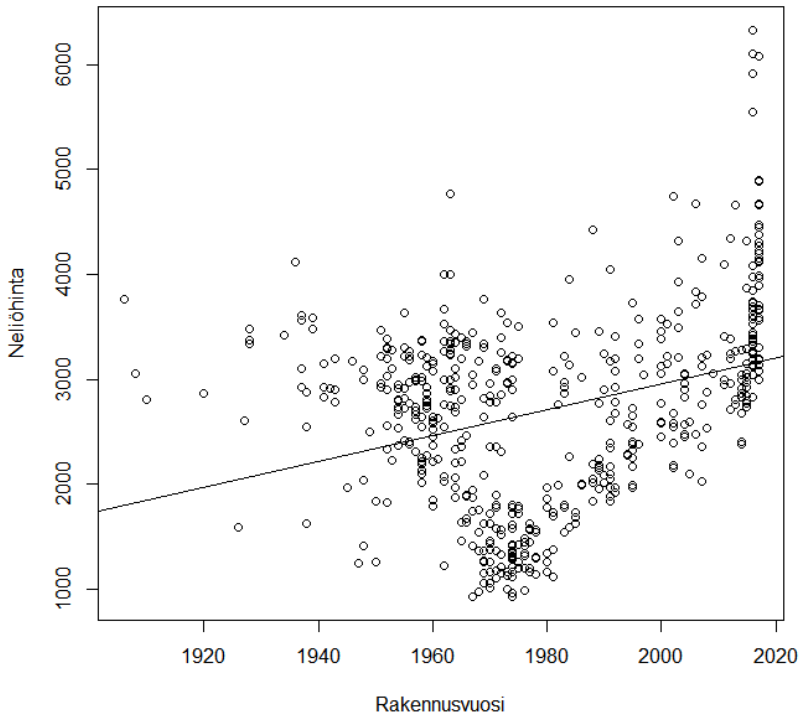
H_0 = parveke ei vaikuta asunnon neliöhintaan

H_1 = parvekkeella on vaikutusta asunnon neliöhintaan.

Sain p-arvoksi 0,0559 > 0,05, jolloin nollahypoteesi ei voida hylätä. Käytetyn datan perusteella ei voi sanoa, että parveke asunnossa vaikuttaisi sen neliöhintaan 5-% merkitsevyystasolla.

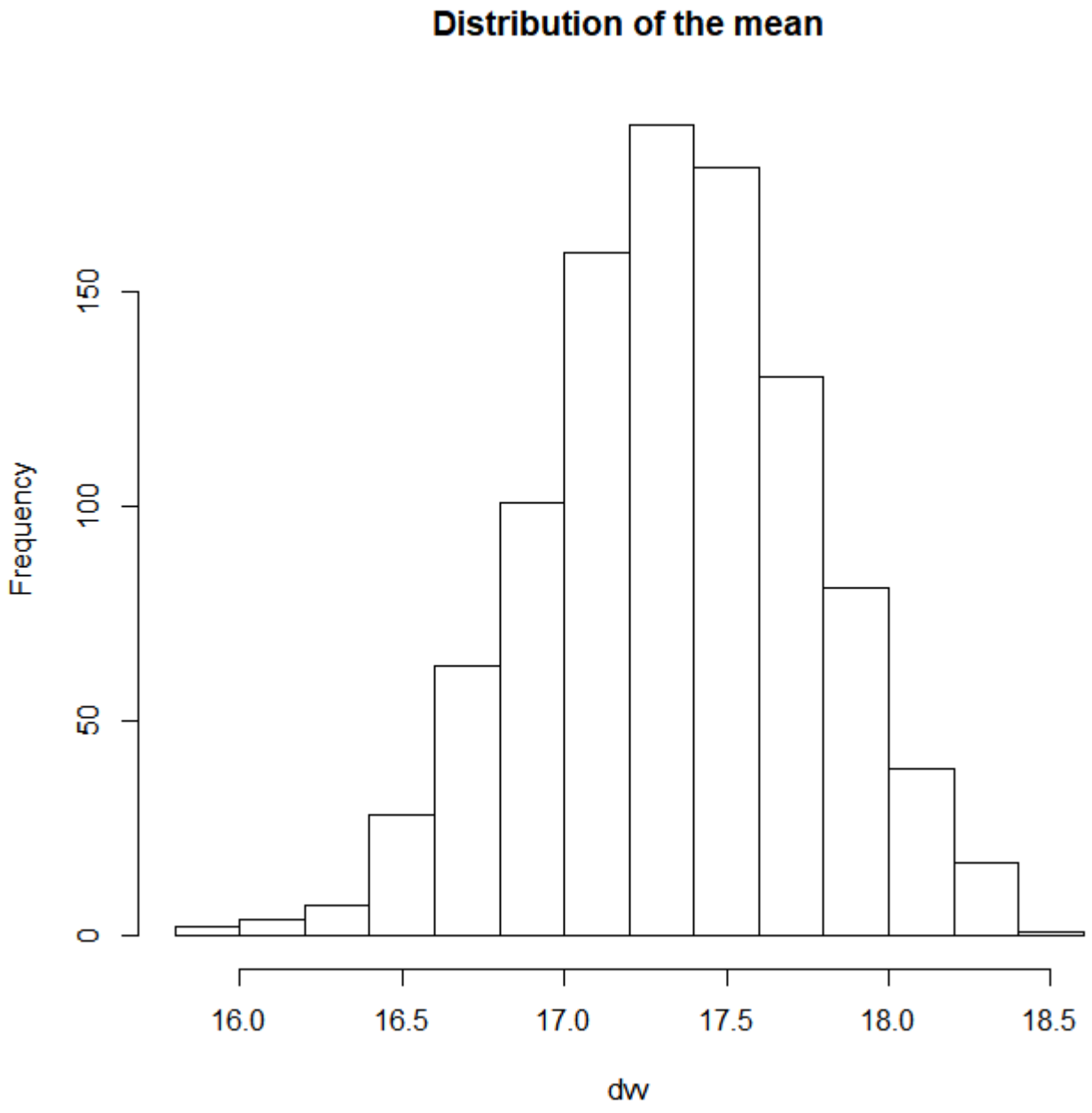
```
> ### RANDOMIZATION TEST ###
>
> Y<-cbind(factor(Parveke), Neliöhinta)
> yl <- which(Y[,1] == 1)
> y2 <- which(Y[,1] == 2)
> Yl <- Y[yl,]; parveke.kyllä <- Yl[,2]
> Y2 <- Y[y2,]; parveke.ei <- Y2[,2]
> parveke.kaikki <- Y[,2]
>
> d <- mean(parveke.kyllä)-mean(parveke.ei)
>
> # Matriisi permutaatioille.
> m<-matrix(nr=178, nc=10000)
>
> # 10000 satunnaista permutaatiota.
> for (i in 1:10000){m[,i]<-sample(parveke.kaikki, 178)}
>
> data1 <- as.data.frame(m[1:89,])
> data2 <- as.data.frame(m[90:178,])
>
> # Molempien keskiarvot.
> means1 <- sapply(data1, mean)
> means2 <- sapply(data2, mean)
> p<-(sum((means1-means2)< -262.4911)+sum((means1-means2)>262.4911))/10000
> p ## 0.0528, eli nollahypoteesia ei voida hylätä.
[1] 0.0559
```

Lopuksi tarkastelin neliöhinnan jakautumista eri vuosina rakennettuihin asuntoihin Monte Carlo -testillä. Alla kuvaaja tilanteen havainnollistamiseksi ja datan perusteella piirretty regressiosuora.



Alustavasti kuvaajaa katsoessa haluaisi sanoa, että rakennusvuosi vaikuttaa asunnon hintaan hyvin suuresti. Tutkitaan, mitä Monte Carlo -testi sanoo asiaan.

Eri havaintojen havaittu etäisyyksien keskiarvo on 10,67. Satunnaisesti muodostettujen otosten havaintojen etäisyyksien keskiarvot hajaantuvat seuraavasti. Keskiarvo on jossain 17.0 ja 18.0 välillä, ja mikään alle 16.0 on jo hyvin epätodennäköinen.



Kuvasta voidaan huomata, että on erittäin epätodennäköistä, että havaintojen etäisyys olisi 10,67, jos neliöhinta eri rakennusvuosina olisi jakautunut simulaation tapaan sattumanvaraisesti.

```
> ### MONTE CARLO ###
>
> X <- cbind(Rakennusvuosi, Neliöhinta); X <- as.data.frame(X)
> plot(X)
> abline(lm(Neliöhinta~Rakennusvuosi))
>
> d <- dist(X); d <- as.matrix(d); d <- as.data.frame(d); diag(d) <- 100000
> dv<-sapply(d,min); dmean<-mean(dv)
>
> dvv<-numeric(1000)
> for (i in 1:1000) {
+     dm<-data.frame(x=runif(564, 1906, 2017),
+     y=runif(564, 923, 6333))
+     dm<-dist(dm); dm<-as.matrix(dm); dm<-as.data.frame(dm);
+     diag(dm)<-100000
+     dvv[i]<-mean(sapply(dm,min)) }
>
> hist(dvv, main="Distribution of the mean"); dmean
[1] 10.67254
```


Lähteet:

aineisto: https://webpages.tuni.fi/uta_statistics/tilasto/tiltp_aineistoja/

Lakennalliset menetelmät ja bayesilaisuuden perusteet -luentomoniste

en.wikipedia.org/wiki/Resampling_(statistics)