Viikko 6

1 Tehtäviä

1. Tutkitaan Pyöräilybarometria lineaarisen regression avulla

   a) Muuta sarakkeiden Aq6 - Aq15 ”En osaa sanoa”-vastaukset puuttuviksi. Luo sen jälkeen summamuuttuja (siis vastausten riveittäiset keskiarvot) sarakkeista Aq6 - Aq15. Muuta summamuuttujan arvot puuttuviksi niillä riveillä, joilla on 5 tai useampi puuttuvaa vastausta. (Tämä on siis täsmälleen sama kuin viikon 5 tehtävä 1. Jos olet tehnyt sen, niin voit käyttää samaa koodia.)

   b) Muodosta lineaarinen malli funktiolla `lm()`, missä selittäjänä on vastaajan ikä ja selitettävänä edellä luotu summamuuttuja. Tallenna malli muuttujaan `fit`.

   c) Piirrä aineistosta kuva, jossa x-akselilla on selittäjä ja y-akselilla selitettävä muuttuja

2. Jatkoa edelliseen tehtävään

   a) Lisää edellisen tehtävän kuvaan regressiosuora valitsemallasi värillä

   b) Tutki komennon `summary(fit)` tulostetta. Tulkitse parametrien estimointituloksia sanallisesti.

   c) Piirrä mallin residuaaleista kuvaaja ja histogrammi. Tulkitse tulosta sanallisesti.

3. Laske edellisen tehtävän lineaarisen mallin parametrien 95% luottamusvälit. Voidaanko hylätä nollahypoteesi H_0: $\beta_0 = 2.0$ (ts. vakio = 2.0) luottamustasolla $1 - \alpha = 0.95$? Vihje: Esimerkki **??**.

4. Tutustu esimerkkiin **??**. Ota nyt selittäjä `x` pois mallista ja piirrä kuva. Vertaa saamaasi mallia esimerkin **??** tulokseen ja pohdi kumpi on parempi selittäjä havainnoille y. Tässä voit myös tutkia funktion `summary()` tulostetta.

5. JTP:n esimerkki 10.1. Tiedetään, että kulhossa 5 palloa. Näistä θ palloa on valkoista ja $5-\theta$ mustaa. Oletetaan, että kaikki valkoisten pallojen määrät $\theta = 0, \ldots, 5$ ovat yhtä todennäköisiä. Nostetaan nyt satunnaisesti 7 palloa (nostettu pallo palautetaan takaisin koriin noston jälkeen, eli kyseessä on otanta takaisinpanolla). Nostetuista palloista 2 on valkoista. Mitkä ovat nyt valkoisten pallojen eri määrien θ todennäköisyydet?

   Tarkastetaan esimerkissä analyyttisesti laskettu todennäköisyysjakauma valkoisten pallojen määrälle simuloimalla. Valkoisten pallojen määrä korissa, eli θ, noudattaa siis diskreettiä tasajakaumaa[1] joukossa $\{0, 1, \ldots, 5\}$, eli $\theta \sim \text{Tas}\{0, 1, \ldots, 5\}$. Jos oletetaan, että korissa on θ valkoista palloa, ja jos merkitään nostettujen valkoisten pallojen määrää satunnaismuuttujalla Y, se noudattaa binomijakaumaa otoskoolla 7 ja onnistumistodennäköisyydellä $\frac{\theta}{5}$, eli $Y|\theta \sim \text{Bin}\left(7, \frac{\theta}{5}\right)$.

[1] Huom. kyseessä on siis diskreetti eikä jatkuva tasajakauma, joten siitä simuloidaan R:ssä käyttäen funktiota `sample` eikä funktiota `rbinom`.

a) Simuloidaan tilannetta 1000000 kertaa. Arvo ensin korissa olevien valkoisten pallojen määrät vektoriin `valkoisia_korissa`. Vihje. `sample`-funktio.

b) Arvo sitten nostettujen valkoisten pallojen määrät 1000000:n pituiseen vektoriin `valkoisia_nostettu` olettaen että jokaisella kerralla korissa on a-kohdassa laskettu määrä valkoisia palloja (esim. jos vektorin `valkoisia_korissa` ensimmäinen arvo on kolme, arvot vektorin `valkoisia_nostettu` ensimmäisen arvon ehdolla $\theta = 3$ ja niin edelleen). Vihje. `rbinom`-funktiolle voi antaa `prob`-argumentiksi vektorin: tällöin R käyttää ensimmäisessä otoksessa todennäköisyytenä vektorin ensimmäistä alkoita, toisessa toista ja niin edelleen.

c) Valitse lopuksi vektoriin `otokset` ne otokset (eli oikeastaan vektorin `valkoisia_korissa` arvot), joilla on nostettu täsmälleen 2 valkoista palloa (eli vektorin `valkoisia_nostettu` arvo on 2). Laske vektorin `otokset` frekvenssitaulu ja normalisoi se jakamalla se niiden otosten, joissa on nostettu 2 valkoista palloa määrällä (eli vektorin `otokset` pituudella). Vertaa simulaatiosi tuottamaa todennäköisyysjakaumaa valkoisten pallojen määrille esimerkissä 10.1 laskettu eksaktiin jakaumaan.

6. JTP:n viikon 5 tehtävä 5: Vuonna 1975 uutisoitiin tutkimuksesta, jonka mukaan 50 % kanadalaisista miehistä käytti värillisiä (= muita kuin valkoisia) alushousuja kun taas amerikkalaisista miehistä sellaisia käytti vain 20 %. Bermudalaisen hotellin asiakaskunta koostui yksinomaan amerikkalaisista ja kanadalaisista siten, että miesasiakkaista 80 % oli amerikkalaisia ja 20 % kanadalaisia. Todennäköisyyslaskentaa opiskellut siivooja huomasi miesasiakkaan huoneessa punaiset alushousut. Millä todennäköisyydellä hän päätteli asiakkaan olevan kanadalaisen?

   Tarkistetaan kynällä ja paperilla laskettu vastaus simuloimalla, eli luomalla otos hotellin asiakkaista ja heidän alushousujen väreistään käyttämällä yllämainittuja osuuksia todennäköisyyksinä.

   a) R:ssä ei ole vakiona funktiota Bernoullin jakauman simuloimiseksi. Bernoullin jakauma on sama asia kuin binomijakauma, jonka otoskoko (huom. R:ssä argumentti `size`) on 1. Luo apufunktio `rbern`, jonka parametrit ovat `n`, eli haluttu otoskoko ja `prob`, eli yksittäisen satunnaiskokeen onnistumisen todennäköisyys. Funktion tulee toimia siis seuraavasti (tietenkin otos on joka kerta eri):

      ```
      > rbern(n=10, prob=0.5)
       [1] 0 1 0 0 0 0 0 0 1 1
      ```

      Vihje: kannattaa hyödyntää valmista `rbinom`-funktiota.

   b) Kirjoita funktio `hotellisimulaatio`, jonka argumenttina on hotellin miespuolisten asukkaiden määrä `n`, ja simuloi edellisen tehtävän tilannetta (käytä yllämainittuja osuuksia todennäköisyyksinä, esim. jos kanadalaisten osuus on 20 %, todennäköisyys että asukas on kanadalainen on 0.2). Funktion tulee palauttaa kanadalaisten osuus niistä asukkaista joilla on värilliset alushousut.

Arvo ensin kanadalaisten määrä, ja sen jälkeen arvo `for`-silmukan ja `if else` - rakenteen avulla niiden asukkaiden, joilla on värilliset alushousut, määrä. Vaikka osaisit tehdä tämän käyttämättä `for`-silmukkaa, käytä sitä kuitenkin tällä kertaa: sama simulaatio tehdään seuraavassa tehtävässä uudelleen hieman fiksummin vektorisoitujen operaatioiden avulla.

Lopuksi laske niiden asukkaiden, jotka ovat kanadalaisia JA joilla on värilliset alushousut, määrä, ja jaa se niiden asukkaiden, joilla ylipäänsä on värilliset alushousut, määrällä.

c) Testaa funktiotasi otoskoolla 100, 10000 ja 1000000. Mikä on simuloinnin perusteella approksimaatio todennäköisyydelle, että punaalushousuinen asiakas on kanadalainen? Jos olet laskenut JTP:n tehtävän käsin, voit verrata simuloimalla laskettua arvoa tarkkaan arvoon.

7. Tehdään edellisen tehtävän funktio hieman ”fiksummin” vektorisoitujen operaatioiden avulla.

   a) Kirjoita funktio `hotellisimulaatio2`, jonka argumentti ja palautusarvo ovat samat kuin edellisessä tehtävässä. Tällä kertaa kuitenkin kanadalaisten määrän arpomisen jälkeen arvo niiden asukkaiden, joilla on värilliset alushousut, määrä käyttämättä `for`-silmukkaa.

   b) Testaa otoskoolla 1000000, että funktiosi antaa samansuuntaisen tuloksen kuin edellisessä tehtävässä.

   c) Vertaa funktioiden `hotellisimulaatio` ja `hotellisimulaatio2` nopeutta otoskoolla 1000000 käyttämällä funktiota `system.time`. Kumpi oli nopeampi, ja oliko ero selkeä (ajat vaihtelevat testikerrasta toiseen, mutta suuruusluokan pitäisi pysyä samana)?

8. Toteutetaan simulaatio vielä kolmannella eri tavalla. Jokaista satunnaiskoetta ei tarvitse arpoa erikseen vektoriin, vaan voidaan arpoa pelkästään kanadalaisten ja värillisalushousuisten määrät suoraan binomijakaumasta.

   a) Toteuta funktio `hotellisimulaatio3`, jonka argumentti ja palautusarvot ovat samat kuin edellisessä tehtävässä. Tällä kertaa kuitenkin arvo ensin kanadalaisten määrä käyttäen suoraan funktiota `rbinom` käyttäen otoskokona (`size`-argumentti) hotellin asukkaiden määrää. Laske tämän jälkeen niiden kanadalaisten määrä, joiden alushousut ovat värilliset, jälleen arpomalla binomijakaumasta, tällä kertaa otoskokona tietenkin kanadalaisten määrä. Sen jälkeen arvo vastaavasti kuinka monella amerikkalaisista on värilliset alushousut, ja laske kanadalaisten värillisalushousuisten osuus kaikista värillisalushousuisista.

   b) Testaa otoskoolla 1000000, että funktiosi antaa samansuuntaisen tuloksen kuin edellisissä tehtävissä.

   c) Vertaa funktioiden `hotellisimulaatio2` ja `hotellisimulaatio3` nopeutta otoskoolla 1000000 käyttämällä funktiota `system.time`. Kumpi oli nopeampi, ja oliko ero selkeä? Jos nopeuseroa on, mistä arvelet sen johtuvan?