Tilastollinen päättely R-ohjelmistolla 2017
Harjoitustyö

Ohjeet

Alla on annettu viisi tehtävää; vastaa kaikkiin näistä. Vastaus koostuu pyydetyistä kuvista, taulukoista ja sanallisista selityksistä. Kokoa kaikkien tehtävien vastaukset yhdeksi raportiksi, jonka voit kirjoittaa haluamallasi tekstinkäsittelyohjelmalla, esim. Wordilla tai LaTeX:illa, ja palauta raportti kurssin Moodle-alueelle .pdf -muodossa **perjantaihin 26.5 klo 23.55** mennessä. Ohjelmakoodia raportissa ei tarvitse olla, mutta osa tuloksista voi olla kätevintä lisätä raporttiin suoraan ohjelman tulosteena.

Palauta lisäksi koodi, jota käytit tehtävien tekemiseen, erillisenä .R-tiedostona. Kaikkien kysyttyjen tulosten täytyy tällä kertaa löytyä myös raportista: jos esimerkiksi kysytään keskiarvoa, kirjoita se näkyviin raporttiin. Tai jos pyydetään piirtämään kuva, liitä se raporttiin. Muista siis raportoida kaikki, mitä tehtävissä kysytään. Palautukseen kuuluu siis **2 tiedostoa**:

1. Raportti **pdf**-tiedostona.
2. R-koodi **.R**-tiedostona

Huomaa: Tehtävämonisteesta on olemassa useampi eri versio. Tehtävät on jaettu Moodlessa opiskelijanumeron mukaan niin, että kaikki eivät saa samaa tehtävämonistetta. Älä siis vastaa kaverin tehtävämonisteen tehtäviin, vaan hae tehtävät kirjautuneena sisään omilla tunnuksillasi!

Harjoitustyön tarkoitus on testata kurssin aikana opittujen asioiden osaamista itsenäisesti, joten harjoitustyön tehtävistä ei voi valitettavasti enää keskustella Moodlessa eikä Presemossa.

Plagiarismi ja opintovilppi, eli esimerkiksi kaverilta saatujen vastausten kopiointi ja niiden palauttaminen omana työnä on harjoitustyössä kiellettyä. Palautettuihin harjoitustöihin voidaan käyttää Urkund-plagiaatintunnistusjärjestelmää ja todettu opintovilppi johtaa kurssisuorituksen hylkäämiseen.

Aineistosta

Aineistona tehtävissä 1-3 on HSL:n asiakastyytyväisyyskysely. Aineistoon on valittu vastaukset ajalta 1.8.2016-17.2.2017. Tehtävässä 4 käsitellään vuorostaan osaa YK:n *sosiaali-indikaattorit*-aineistosta. Molemmat aineistot ovat saatavilla kurssin Moodle-alueella.

Tehtävät

1. Lista, summamuuttuja ja yhden muuttujan jakaumien tarkastelua.

    a) Kirjoita funktio `yhteenveto`, joka ottaa argumentikseen tarkasteltavan muuttujan. Funktion tulee palauttaa lista, jossa on seuraavat neljä komponenttia:

    - argumenttina annetun muuttujan frekvenssitaulu

- argumenttina annetun muuttujan ei-puuttuvien arvojen määrä ja puuttuvien määrä vektorina siten, että vektorin ensimmäisessä alkiossa on ei-puuttuvien määrä ja toisessa puuttuvien määrä (anna komponentin nimeksi `n_ja_puuttuvat`)
- argumenttina annetun muuttujan keskiarvo (anna komponentin nimeksi `keskiarvo`)
- argumenttina annetun muuttujan mediaani (anna komponentin nimeksi `mediaani`)

Varmista, että keskiarvoon ja mediaaniin ei lasketa puuttuvia arvoja mukaan, jos sellaisia on tarkasteltavassa muuttujassa.

b) Kirjoita funktio `summamuuttuja`, joka ottaa argumenteikseen:

- tarkasteltavien muuttujien nimet merkkijonovektorina
- aineiston taulukkona eli `data.frame`:na
- suurimman sallittujen puuttuvien arvojen määrän / vastaaja (aseta argumentin oletusarvoksi tarkasteltavien muuttujien määrä - 1, eli ensimmäisenä argumenttina annetun vektorin pituus - 1).

Funktion tulee palauttaa summamuuttuja tarkasteltavista muuttujista, eli vektori johon on laskettu jokaiselle vastaajalle hänen vastaustensa keskiarvo tarkasteltavista muuttujista. Käytä keskiarvoja laskiessasi argumenttia `na.rm = TRUE`, mutta muuta summamuuttujan arvot niiden vastaajien, joilta puuttuu vastaus useampaan kysymykseen kuin mikä on argumentiksi annettu suurin sallittu puuttuvien vastausten määrä, kohdalta puuttuviksi.

c) Lataa HSL:n asiakastyytyväisyyskyselyn aineisto taulukkoon `asty`, ja laske funktiosi `summamuuttuja` avulla vastaajien keskimääräinen tyytyväisyys kuljettajien toimintaan asteikolla 1-5 muuttujaan `tyyt_kulj` laskemalla vastaajakohtaiset keskiarvot muuttujista `K1A1`, `K1A2` ja `K1A3`. Määrää funktiosi argumentilla summamuuttujan arvo puuttuvaksi, jos vastaajalta puuttuu vastaus useampaan kuin yhteen kysymykseen. Funktiokutsun tulisi siis toimia esimerkiksi seuraavasti:

```
asty$tyyt_kulj <- summamuuttuja(c('K1A1', 'K1A2', 'K1A3'),
data = asty, max_puuttuvat = 1)
```

Selvitä `yhteenveto`-funktiollasi summamuuttujan `tyyt_kulj` keskiarvo ja mediaani sekä niiden vastaajien lukumäärä, joille on laskettu arvo summamuuttujaan. Miltä vastaajien keskimääräinen tyyväisyys kuljettajien toimintaan vaikuttaa suurin piirtein keskiarvon ja mediaanin perusteella? Muuttujien asteikot ja niiden kuvaukset löytyvät koodikirjasta. Piirrä summamuuttujan jakaumasta oranssi histogrammi ja anna sille kuvaava otsikko. Selvitä `yhteenveto`-funktiosi tulostaman frekvenssitaulun avulla summamuuttujan yleisin arvo. Miten se näkyy histogrammissa? Piirrä vielä pystysuora viiva histogrammiin summamuuttujan mediaanin kohdalle ja tee viivasta riittävän paksu (esim. argumentilla `lwd=5`).

d) Laske funktiosi `summamuuttuja` avulla vastaajien keskimääräistä tyytyväisyyttä HSL:n palveluihin kuvaava summamuuttuja `tyyt_hsl` laskemalla vastaajakohtaiset keskiarvot muuttujista `K1A4`, `K1A5`, `K1A6`,

`K2A2`, `K2A3`, `K2A4`, `K2A5` ja `K2A6`. Määrää funktiosi argumentilla summamuuttujan arvo puuttuvaksi, jos vastaajalta puuttuu vastaus useampaan kuin viiteen kysymykseen. Tarkastele `yhteenveto`-funktiollasi summamuuttujan keskiarvoa ja mediaania. Mitä voit näiden perusteella sanoa vastaajien yleisestä tyytyväisyydestä HSL:n palveluihin suurin piirtein?

2. Osa-aineistoja ja tilastollista hypoteesintestausta

   a) Tee osa-aineisto, johon kuuluvat vain vastaajat, jotka ovat vastanneet kyselyyn joko linjalla 550 tai 560. Rajaa aineistoa siis muuttujan `LINJA` avulla ja tallenna osa-aineisto taulukkoon `runkolinjat`. Muunna osa-aineiston muuttuja `LINJA` nyt faktoriksi, jonka tasot ovat `Jokerivanha` ja `Jokeriuusi`. Vanhalla jokerilinjalla tarkoitetaan linjaa 550 ja uudella linjaa 560.

   b) Ristiintaulukoi faktoriksi muuttamasi muuttuja `LINJA` osa-aineiston muuttujan `K1A3` (Kuljettajien ajotapa on miellyttävä ja sujuva) kanssa. Laske taulukosta joko rivi- tai sarakeprosentit sen mukaan, kummalla niistä on järkevämpi tulkinta. Voit esittää luvut desimaalilukuina tai prosenttimuodossa, mutta pyöristä luvut sopivaan esitystarkkuuteen. Onko linjojen välillä silmämääräisesti havaittavissa merkittävää eroa mielipiteessä kuljettajien ajotapaan näiden prosenttien perusteella?

   c) Testaa osa-aineistossasi kahden otoksen t-testillä [1] nollahypoteesia

   $$H_0 : \mu_{550} = \mu_{560}$$

   kaksisuuntaista vastahypoteesia

   $$H_1 : \mu_{550} \neq \mu_{560}$$

   vastaan eli sitä kokevatko runkolinjojen 550 ja 560 matkustajat keskimääräisen kuljettajien ajotavan yhtä miellyttäväksi ja sujuvaksi. Käytä merkitsevyystasoa $\alpha = 0.1$. Mikä on testin tulos? Onko kahden eri runkolinjan matkustajien keskimääräisissä mielipiteissä kuljettajien ajotavasta eroa keskenään?

   d) Kuvitellaan tilanne, että kaikilla HSL-alueen bussilinjoilla mielipiteiden keskiarvo kuljettajien ajotapaa käsittelevässä muuttujassa `K1A3` olisi pitkän aikaa ollut tasolla 4.0. Testaa osa-aineistossasi yhden otoksen t-testillä nollahypoteesia

   $$H_0 : \mu_{550} = 4.0$$

   kaksisuuntaista vastahypoteesia

   $$H_1 : \mu_{550} \neq 4.0$$

   vastaan eli sitä onko runkolinjan 550 matkustajien mielipiteiden todellinen keskiarvo muuttujassa `K1A3` tasolla 4.0. Käytä merkitsevyystasoa $\alpha = 0.05$. Mikä on testin tulos ja sen mukainen johtopäätös?

[1] Ryhmien otoskoot ovat suurehkoja, jolloin ryhmien otoskeskiarvojen voidaan suurin piirtein olettaa noudattavan normaalijakaumaa, mikä mahdollistaa t-testin käyttämisen.

3. Päivämäärien käsittelyä ja luottamusvälien visualisointia

   a) Palataan takaisin käsittelemään koko aineistoa eli taulukkoa `asty`. Aineisto on etukäteen rajattu siten, että vastauksia on vain vuosilta 2016 ja 2017. Vastausten päivämäärät löytyvät muuttujasta `PAIVAMAARA` muodossa "vvvv-kk-pp". Määritellään nyt vastaajan ikä vähentämällä vastaajan syntymävuosi vuosiluvusta, jolloin tämä on vastannut kyselyyn.[2] Määritä ensin aineistoon uusi numeerinen muuttuja `vuosi`, joka kertoo, minä vuonna kyselyyn vastattiin. Määritä sitten aineistoon uusi muuttuja `ika` vähentämällä vuosiluvusta vastaajan syntymävuosi eli muuttuja `T7`.

   b) Tee funktio `KeskiarvoJaVali`, joka ottaa argumentikseen vektorin lukuja ja halutun luottamustason, ja palauttaa toisen vektorin joka sisältää funktiolle syötettyjen lukujen keskiarvon, sekä niitä vastaavan t-luottamusvälin ala- ja ylärajan.

   c) Luokittele vastaajien iät viiteen luokkaan: Alle 18-vuotiaat, 18-34-vuotiaat, 35-64-vuotiaat, 65-74-vuotiaat ja vähintään 75-vuotiaat. Tallenna ikäluokat aineistoon uuteen muuttujaan `ikaluokka`. Tee sitten osa-aineisto, johon valitset vain bussilinjalla 72 kyselyyn vastanneet. Rajaa aineistoa siis muuttujan `LINJA` avulla. Muuttujan kuvaukset löytyvät koodikirjasta. Tallenna osa-aineisto taulukkoon `asty_72`. Laske osa-aineistosta funktiota `KeskiarvoJaVali` käyttäen muuttujan `K1A3` (Kuljettajien ajotapa on miellyttävä ja sujuva) keskiarvo ja 95% t-luottamusväli jokaiselle ikäryhmälle.

   d) Visualisoi muuttujan `K1A3` (Kuljettajien ajotapa on miellyttävä ja sujuva) vaihtelua ikäryhmittäin bussilinjalla 72 kyselyyn vastanneiden keskuudessa. Piirrä ensin tyhjä kuva ja rajoita sen x-akseli välille 1-5 (ikäryhmien lukumäärä). Käytä kuitenkin argumenttia `xaxt="n"`, jolloin x-akselin pisteille ei tule mitään kuvauksia, koska haluamme numeroiden 1-5 sijasta x-akselin pisteille kuvauksiksi ikäryhmien tasot. Rajoita y-akseli välille 1-5 (`K1A3`-muuttujan vaihteluväli). Anna x-akselin otsikoksi ”Ikäryhmä” ja y-akselin otsikoksi "Mielipide kuljettajien ajotavasta". Piirrä sitten kuvaan jokaiselle viidelle ikäryhmälle lasketut muuttujan `K1A3` keskiarvot pisteinä. Käytä pisteenä mieleistäsi merkkiä (argumentti `pch`). Piirrä jokaisen ikäryhmän pisteen ympärille näiden luottamusvälejä vastaavat (pystysuorat) janat.

   Selvitä, kuinka saat x-akselin pisteiden kuvauksiksi ikäryhmien tasot (alle 18, 18-35, jne.). Tätä ei ole selitetty kurssimateriaalissa, joten joudut kaivamaan tiedon muualta. Vihje: funktio `axis`.

   Arvioi silmämääräisesti tai katso edellisestä kohdasta, millä kahdella ikäryhmällä luottamusvälit ovat lyhimmät. Pohdi sanallisesti, mistä erot ikäryhmien luottamusvälien pituuksissa johtuvat.

4. Muuttujien tutkimista lineaarisella mallilla

   a) Lataa YK:n *sosiaali-indikaattorit*-aineisto taulukkoon `yk` ja lisää siihen uusi muuttuja `logBKT` ottamalla luonnollinen logaritmi asukaskohtaista bruttokansantuotetta kuvaavasta muuttujasta `BKT`. Sovi-

[2] Tämä ei tuota kaikille oikeaa ikää, mutta heittää enintään vuoden.

ta sitten aineistolla lineaarinen malli, jossa selität vuotuista väestönkasvua (%) `vaestonkasvu` muuttujan `logBKT` avulla. Tarkastele `summary`-funktion mallille palauttamaa tulostetta. Millainen suhde väestönkasvulla ja bruttokansantuotteen logaritmilla vaikuttaa tämän perusteella olevan?

b) Visualisoi muuttujien `logBKT` ja `vaestonkasvu` välistä suhdetta piirtämällä niistä hajontakuva. Erota kuvassa eri alueisiin kuuluvat valtiot määrittämällä pisteiden värit muuttujan `alue` mukaan ja piirrä kuvaan regressiosuora.

Etsi hajontakuvasta kaksi havaintoa, jotka poikkeavat mielestäsi eniten regressiosuorasta. Merkitse kuvaan niitä vastaavien valtioiden nimet `text`-funktion avulla.

c) Lisää `lukutaito_naiset` lineaariseen malliin toiseksi selittäjäksi. Tarkastele jälleen `summary`-funktion tulostetta, ja laske sen lisäksi 99% luottamusvälit mallin kertoimille. Millainen vaikutus muuttujan `lukutaito_naiset` lisäämisellä oli malliin? Mistä arvelisit tämän tuloksen johtuvan?

5. Bayes-päättelyä simuloimalla

Pekka on saanut Maijalta kolikon, joka palauttaa heitettynä kruunia eräällä tuntemattomalla todennäköisyydellä $0 < \theta < 1$. Pekka haluaa selvittää, mikä parametrin θ oikea arvo todennäköisesti on, joten hän päättää yrittää määrittää sitä kokeellisesti heittämällä saamaansa kolikkoa muutamia kertoja. Seitsemällä heitolla Pekka havaitsi yhteensä viisi kruunaa.

Tässä havaittujen kruunien lukumäärä Y noudattaa binomijakaumaa otoskoolla 7 ja onnistumistodennäköisyydellä θ, eli $Y|\theta \sim Bin(7, \theta)$. Pekalla ei ole mitään esitietoa parametrin θ arvosta, joten hän päättää olettaa sen kaikkien mahdollisten arvojen olevan yksinkertaisesti yhtä todennäköisiä. Toisin sanoen θ:n ajatellaan siis olevan tasajakautunut välillä $(0, 1)$, eli $\theta \sim U(0, 1)$. Tutkitaan tilannetta simuloimalla, mitä parametrista θ voidaan päätellä saadun tuloksen perusteella.

a) Kirjoita funktio, joka simuloi tehtävän tilannetta. Funktion tulee ottaa argumenttinaan simulaation otoskoko n ja palauttaa vektori kaikista generoiduista θ:n arvoista, joilla arvottu kruunien lukumäärä on tasan viisi.

Toteuta funktio generoimalla ensiksi tasajakaumasta n kappaletta mahdollisia onnistumistodennäköisyyksiä. Arvo sitten näillä saatujen kruunien määrät, kun kolikkoa heitetään seitsemän kertaa jokaisella n onnistumistodennäköisyydellä. Valitse lopuksi generoiduista onnistumistodennäköisyyksistä ne, joilla saatu kruunien lukumäärä vastaa Pekan saamaa havaintoa $Y = 5$.

Suorita valmis funktio kertaalleen simulaation otoskoolla $n = 100000$, piirrä sen palauttamista arvoista histogrammi ja laske saaduista arvoista lisäksi tavallisia tunnuslukuja. Mitä voit sanoa parametrin θ jakaumasta simulaation avulla saadun approksimaation perusteella?

b) Maija seuraa sivusta Pekan koetta. Hän tietää, että hänen Pekalle antamansa lantti on yksi viidestä samannäköisestä kolikosta, joilla on

tunnetut onnistumistodennäköisyydet $\theta_1 = 0.2$, $\theta_2 = 0.4$, $\theta_3 = 0.5$, $\theta_4 = 0.7$ ja $\theta_5 = 0.8$. Maija valitsi kolikon sattumanvaraisesti eikä siksi tiedä, mitä näistä kolikosta Pekka heittää. Pekan tavoin hän päättää täten olettaa, että kaikki viisi kolikkoa ovat yhtä todennäköisiä.

Asetelmalta tilanne eroaa nyt a)-kohdasta esitiedon osalta. Tällä kertaa tutkittavan parametrin θ tiedetään noudattavan kolikkojen joukon $\{1, 2, 3, 4, 5\}$ diskreettiä tasajakaumaa.

Tee taas funktio, joka simuloi koetta hyödyntäen nyt uutta esitietoa. Funktion on otettava jälleen argumentikseen simulaation otoskoko n. Tällä kertaa sen on kuitenkin palautettava havaitut todennäköisyydet kullekkin eri kolikolle esimerkiksi tauluna tai nimettynä vektorina.

Toteuta funktio arpomalla aluksi n kappaletta kolikoita, ja generoi niitä vastaavilla onnistumistodennäköisyyksillä kruunien määrät n:lle seitsemällä kolikon heitolle. Rajaa arvotuista kolikoista ne, joilla generoitu kruunien määrä vastaa Pekan havaintoa $Y = 5$, ja laske kunkin mahdollisen kolikon 1-5 osuudet tässä osajoukossa.

Suorita valmis funktio jälleen otoskoolla $n = 100000$. Mitä näistä viidestä kolikosta Pekka todennäköisimmin heittelee?