Data-analyysi R-ohjelmistolla, kevät 2015

Viikko 4

Tehtäviä

1. Luodaan vastaajien tyytyväisyyttä Helsingin pyöräilyolosuhteisiin kuvaava summamuuttuja.

   a) Valitse osa-aineisto, joka sisältää muuttujat `Aq6-Aq15`, jotka kuvaavat vastaajien tyytyväisyyttä erilaisiin pyöräilyä koskeviin asioihin Helsingissä, ja tallenna se taulukkoon. Seuraavien kohtien tehtävät on helpompi tehdä tätä aputaulukkoa käyttäen, jolloin muuttujia ei tarvitse joka kerta valita erikseen. Vihje: Esimerkki 5.

   b) Tulosta muuttujien `Aq6-Aq15` frekvenssitaulut. Vihje: `apply()`.

   c) Muuta muuttujien `Aq6-Aq15` "En osaa sanoa"-vastaukset puuttuviksi. Tarkista onnistuminen tulostamalla muuttujien frekvenssitaulut uudelleen.

   d) Laske vastausten `Aq6-Aq15` keskiarvot *riveittäin* (siis jokaiselle riville lasketaan vastaajan vastausten keskiarvo kysymyksistä `Aq6-Aq15`) ja tallenna tulos alkuperäiseen taulukkoon `pb` uudeksi muuttujaksi `tyytyvaisyys`. Vihje: `apply()` ja `cbind()`.

2. Aineistossa on valmis viiteen luokkaan luokiteltu ikä-muuttuja `Aikalk1`. Luodaan vastaava muuttuja itse.

   a) Luo aineistoon uusi muuttuja `ika5`, jonka luokittelee vastaajat viiteen ikäryhmään *täsmälleen* samalla tavalla kuin `Aikalk1`. Vihje: `cut()` ja `cbind()`.

   b) Testaa onnistuminen ristiintaulukoimalla luomasti `ika5` ja `Aikalk1` (käytä `useNA`-argumenttia, niin näet myös mahdolliset puuttuvat arvot). Tuloksena tulisi olla taulu, jossa kaikki muut paitsi diagonaalialkiot ovat nollia. Jos epäonnistuit, palaa a-kohtaan ja yritä uudelleen.

   c) Nimeä muuttujan `ika5` tasot uudelleen jotta saat siistimmän tulosteen, esimerkiksi samoiksi kuin muuttujan `AikAlk1` tasot. Tulosta muuttujan `ika5` frekvenssitaulu

   d) Laske `tyytyvaisyys`-summamuuttujan keskiarvot ikäluokittain. Vihje: Esimerkki 9.

3. Luottamusvälin laskeminen.

   a) Laske kaksisuuntainen t-luottamusväli muuttujan `tyytyvaisyys` keskiarvolle 95%:n luottamustasolla sijoittamalla oikeat arvot luottamusvälin kaavaan

   $$\left[\bar{y} - t_{n-1}(0.025)\frac{s}{\sqrt{n}}, \quad \bar{y} + t_{n-1}(0.025)\frac{s}{\sqrt{n}}\right],$$

   missä $\bar{y}$ on muuttujan `tyytyväisyys` keskiarvo, s sen keskihajonta ja n otoskoko (huom. ei sama kuin koko aineiston otoskoko, koska

muuttujalla `tyytyväisyys` puuttuvia arvoja). Merkintä $t_{n-1}(0.025)$ tarkoittaa vapausasteen $n-1$ t-jakauman 0.025-yläkvantiilia, eli pistettä, jonka oikealla puolella on 2.5% jakauman todennäköisyysmassasta, eli vasemmalla puolella on 97,5% jakauman todennäköisyysmassasta, eli kyseessä on 0.975-kvantiili. Vihje `qt()`.

b) Tarkista tulos komennolla `t.test(pb$tyytyvaisyys)$conf.int`.

4. Tarkastellaan kysymyksen Aq3 (tyytyväisyys Helsinkiin pyöräilykaupunkina) vastauksia sukupuolittain. Laske 95% kaksisuuntaiset t-luottamusvälit miehille ja naisille erikseen. Vertaa luottamusvälien pituuksia toisiinsa.

5. Pohditaan seuraavaksi kurssin Johdatus Tilastolliseen päättelyyn harjoitusten 3 tehtävää 2. Tehtävässä tutkitaan auton polttoaineen keskikulutusta kaupunkiajossa. Ajoa suorittaa kuusi eri kuljettajaa, joiden keskikulutukseksi havaitaan: 6.1, 5.7, 5.9, 6.8, 6.7 ja 6.0.

   a) Tallenna annetut mittaustulokset vektoriin ja laske sen keskiarvo ja keskihajonta.

   b) Laske keskikulutukselle 95% luottamusväli käyttäen funktiota `t.test()`.

6. Tutkitaan nyt tapausta jossa $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ ovat riippumattomia. Parametrin μ suurimman uskottavuuden estimaattori on tunnetusti satunnaismuuttujien Y_i, $i = 1, \ldots, n$ keskiarvo, siis $\hat{\mu} = \overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

   Tutkitaan nyt estimaattorin $\hat{\mu}$ kaksisuuntaisen 95% t-luottamusvälin merkitystä, kun molemmat parametrinarvot ovat tuntemattomia. Käytetään simuloinnissa jakaumaa N(0,1).

   a) Simuloi 200 otosta joissa jokaisessa otoskoko $n = 100$, laske otoksista kaksisuuntaiset 95% t-luottamusvälit odotusarvolle ja tallenna ne muuttujaan `luottamus`. (Vihje: `sapply()` ja esimerkki 10). *Voit myös halutessasi laskea ylä- ja alarajat omiin vektoreihinsa.*

   b) Luo tyhjä `plot()`-ikkuna, jossa y-akseli on välillä $[-1, 1]$ ja x-akseli välillä $[0, 200]$. Vihje: Esimerkki 12.

   c) Piirrä kuvaan pystyviivat luottamusvälin ylärajasta ja alarajaan käyttäen funktiota `segments()`. Vihje: Esimerkki 15.

   d) Piirrä kuvaan punainen viiva tasolle y=0. Vihje: `abline()`.

   e) Tarkastele silmämääräisesti: Peittävätkö luottamusvälit todellisen parametrinarvon noin noin 95%:ssa tapauksista?

7. Tarkastellaan tehtävässä 6 piirrettyjä luottamusvälejä, jotka tallennettiin muuttujaan `luottamus` (*tai ylä- ja alarajat omiin vektoreihinsa*). Laske kuinka suuri osuus luottamusväleistä peittää parametrin todellisen arvon $\mu = 0$.

8. Piirrä tehtävän 6 kuva uudestaan, nyt käyttäen ensin otoskokoa $n = 50$ ja sitten $n = 500$. Mitä huomaat?