Data-analyysi R-ohjelmistolla, kevät 2015

Viikko 5

1 Tehtäviä

1.1 Tehtäviä

1. Puuttuvien arvojen huomioon ottaminen summamuuttujaa laskettaessa.

   a) Lataa Pyöräilybarometri-aineisto ja valitse osa-aineisto muuttujista `Aq6 - Aq15` samalla tavalla kuin viime viikon tehtävässä 1. Muuta "En osaa sanoa"-vastaukset puuttuviksi.

   b) Laske kuinka monta puuttuvaa vastausta kullakin vastaajalla on yhteensä kysymyksiin `Aq6 - Aq15` (esimerkiksi jos vastaaja on jättänyt vastaamatta kolmeen näistä kymmennestä kysymyksestä, muuttujan arvoksi tulee tämän vastaajan kohdalla 3) ja tallenna tulos vektoriin `puuttuvat`

   c) Tulosta muuttujan `puuttuvat` frekvenssitaulu ja laske kuinka moni vastaaja on jättänyt vastaamatta viiteen tai useampaan kysymykseen. Vihje: yksiulotteinen taulu on nimetty vektori, joten voit summata sen alkioita. Taulun kaikkien alkioiden summan saa laskettua komennolla `sum(taulu[1:length(taulu)])`, jos taulu on tallennettu muuttujaan `taulu`. Ole kuitenkin tarkkana indeksien kanssa!

   d) Luo summamuuttuja `tyytyvaisyys` samalla tavalla kuin viime viikon tehtävässä yksi. Käytä tällä kertaa argumenttia `na.rm=TRUE`, jolloin muuttujan arvo lasketaan, jos vastaaja on vastannut yhteenkin kysymyksistä. Tallenna summamuuttuja sarakkeeksi alkuperäiseen taulukkoon `pb`. Laske kuinka monta puuttuvaa arvoa summamuuttuja saa.

   e) Vastaajien, jotka ovat jättäneet vastaamatta useaan kysymyksistä joista summamuuttuja lasketaan, vastauksia ei yleensä haluta ottaa huomioon summamuuttujaa laskettaessa. Muuta siis summamuuttujan `tyytyvaisyys` arvo puuttuvaksi, jos vastaaja on jättänyt vastaamatta viiteen tai useampaan kysymykseen (eli muuttujan `puuttuvat` arvo on 5 tai suurempi. Laske kuinka monta puuttuvaa arvoa summamuuttuja `tyytyvaisyys` nyt saa, ja tarkasta että tulos on sama kuin c-kohdassa laskettu arvo.

2. Testataan poikkeaako edellisessä tehtävässä laskettu `tyytyvaisyys-muuttujan` keskiarvo tilastollisesti merkitsevästi 2.2:sta merkitsevyystasolla 0.05. Testataan siis t-testillä nollahypoteesia $H_0 : \mu = 2.2$ kaksisuuntaista vastahypoteesiä $H_1 : \mu \neq 2.2$ vastaan.

   a) Lasketaan ensin testisuure käsin. Laske t-testisuureen arvo kaavasta

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}},$$

missä $\overline{y}$ on muuttujan `tyytyvaisyys` otoskeskiarvo, s sen otoskeskihajonta ja n sen otoskoko (huom. n on todellinen otoskoko, eli puuttuvia arvoja ei lasketa mukaan) ja μ_0 nollahypoteesin mukainen parametrin arvo.

b) Laske testin p-arvo kaavasta

$$p = P(|T| \geq |t|) = 2(1 - F_{n-1}(|t|)),$$

missä F_{n-1} on t:n jakauman vapausasteella $n-1$ kertymäfunktio, joka saadaan R:ssa funktiolla `pt`.

c) Tarkasta laskemasi t-testisuureen arvo ja p-arvo funktiolla `t.test`. Tulkitse testin tulosta.

3. Tarkastellaan edellisen tehtävän tulosta visuaalisesti.

a) Piirrä t:n jakauman vapausasteella $n-1$ (missä n on edellisen tehtävän `tyytvaisyys`-muuttujan otoskoko) jakauman tiheysfunktion kuvaaja välillä $[-4, 4]$. T:n jakauman tiheysfunktion arvot saadaan R:ssä funktiolla `dt`.

b) Liitä kuvaan pystysuorat viivat kohtiin $t_{n-1}(0.025)$ ja $t_{n-1}(0.975)$, eli t:n jakauman vapausasteella $n-1$ 2.5 prosentin ja 97.5:n prosentin kvantiilien kohtiin (vihje: funktio `qt`). Tee viivoista haluamasi väriset.

c) Liitä kuvaan pystysuora viiva (erivärinen kuin edelliset viivat) edellisen tehtävän t-testisuureen arvon kohtaan.

d) Miten tulkitset kuvaa: missä ovat merkitsevyystason 0.05 t-testin kriittiset alueet, ja sijoittuuko t-testisuureen arvo kriittiselle alueelle?

4. Jatkoa edelliseen tehtävään.

a) Simuloi sata t-testisuureen arvoa, eli 100:n kokoinen otos t:n jakaumasta vapausasteella $n-1$, missä n on edelleen `tyytyvaisyys`-muuttujan otoskoko.

b) Lisää edellisen tehtävän kuvaan jälleen uudella värillä pystysuorat viivat a-kohdassa simuloitujen arvojen kohtiin.

c) Arvioi silmämääräisesti kuinka moni viivoista, eli simuloiduista t-testisuureen arvoista, on testin kriittisellä alueella? Entä onko yksikään simuloiduista t-testisuureen arvoista yhtä kaukana jakauman hännässä kuin havaittu (tehtävässä kolme laskettu) t-testisuureen arvo?

d) Tarkasta edellisen kohdan arvio laskemalla kuinka moni a-kohdassa simuloiduista testisuureen arvoista on testin kriittisellä alueella, eli pienempää kuin $t_{n-1}(0.975)$ tai suurempaa kuin $t_{n-1}(0.025)$ (huomaa että kvantiileille on tässä käytetty JTP:n merkintöjä, jotka ovat "väärinpäin": piste $t_{n-1}(0.975)$ on t:n jakauman 2.5 prosentin kvantiili ja piste $t_{n-1}(0.025)$ 97.5 prosentin kvantiili)?

5. Lisää simulointia

a) Simuloi tällä kertaa sata n:n kokoista (n on `tyytyvaisyys`-muuttujan otoskoko) otosta normaalijakaumasta, jonka keskiarvo on 2.2, keskihajonta `tyytyvaisyys`-muuttujan keskihajonta. Vihje: `replicate`.

b) Laske jokaiselle otokselle muuttuja z kaavasta

$$z = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}},$$

missä $\overline{y}$ on kunkin otoksen otoskeskiarvo, $\mu = 2.2$, σ on `tyytyvaisyys`-muuttujan otoskeskihajonta, jota käytettiin simuloitaessa jakauman parametrina ja n on otoksien koko (eli `tyytyvaisyys`-muuttujan otoskoko). Tallenna tulokset vektoriin `z`. Vihje: `apply` (Voit myös tehdä a- ja b-kohdan saman `replicate`-lauseen sisällä viime viikon tehtävän 6 a) tyyliin, jos uskallat).

c) Piirrä standardinormaalijakauman tiheysfunktion kuvaaja välillä $[-4, 4]$.

d) Liitä kuvaan pystysuorat viivat kohtiin $z_{0.975}$ ja $z_{0.025}$, eli standardinormaalijakauman 2.5 prosentin ja 97.5 prosentin kvantiilien kohdalle mieleiselläsi värillä.

e) Lisää kuvaan pystysuorat viivat vektorin `z` arvojen kohtiin jollain toisella värillä.

f) Vertaa lopputulosta edellisen tehtävän kuvaan, ja selitä mitä havaitset (tätä ei tarvitse tietää, mutta mieti asiaa ja selitä näkemyksesi)?

6. Lista funktion palautusarvona.

a) Kirjoita funktio joka laskee kaksisuuntaisen yhden otoksen t-testin t-testisuureen arvon ja p-arvon (tehtävän 2 kaavojen avulla, käyttämättä t.test-funktiota). Funktiolle annetaan parametreina tarkasteltavan muuutujan otoskeskiarvo $\overline{y}$, otoskeskihajonta s, otoskoko n ja nollahypoteesin mukainen parametrin arvo μ_0.

Funktion tulee palauttaa lista. jonka ensimmäinen komponentti on `t_testisuure`, joka sisältää t-testisuureen arvon, toinen komponentti `p_arvo`, joka sisältää testin p-arvon ja kolmas komponentti on `luottamusvali_95`, joka on vektori joka sisältää keskiarvon 95 prosentin luottamusvälin ala- ja ylärajan (laske tämäkin käyttämättä `t.test`-funktiota, kaava löytyy viime viikon harjoitusten tehtävästä kolme).

b) Testaa että funktiosi toimii oikein syöttämällä siihen tehtävän 2 arvot ja tarkastamalla että saat samat tulokset.

7. JTP:n harjoitusten 4 tehtävä 4 a). Testaa t-testillä edellisessä tehtävässä luomaasi funktiota käyttäen nollahypoteesia $H_0 : \mu = 3$ kaksisuuntaista vastahypoteesia $H_1 : \mu \neq 3$ vastaan merkitsevyystasolla 0.05, kun

a)
$$n = 10, \quad \overline{y} = 3.952, \quad s^2 = 1.981$$

b)
$$n = 15, \quad \overline{y} = 3.411, \quad s^2 = 1.321,$$

ja tulkitse testien tulokset sanallisesti. Huomaa, että tehtävässä on annettu otosvarianssit s^2, eikä otoskeskihajontoja s.

8. Tutkitaan esimerkin 23 tapausta. Muuttuisiko päättelyn tulos, jos Hiivalle 2 olisikin havaittu 140 kohonnutta ja 60 kohoamatonta taikinaa?

9. Esimerkin 23 fiktiivinen leipomo päättää tehdä uuden aluevaltauksen ja värjätä taikinansa punaiseksi, vihreäksi ja siniseksi. Jokaista leipälajia valmistetaan päivittäin, värikohtaiset määrät saattavat vaihdella. Värikohtainen leivottujen ja myytyjen leipien lukumäärä kirjataan ylös.

   Tutki riippumattomuustestin avulla nollahypoteesia, jonka mukaan leivän väri ei vaikuta sen myyntiin. Viikon tuotanto ja myynti ovat taulukoituna alla. *Vihje: Pohdi ensin mitä tässä taulukossa on ja mitä pitää vielä laskea. Ehkä kannattaa taulukoida väreittäin viikon aikana myydyt ja myymättömät leivät.*

	Punainen		Vihreä		Sininen	
	Leivottu	Myyty	Leivottu	Myyty	Leivottu	Myyty
Maanantai	15	12	23	20	4	4
Tiistai	24	10	13	8	16	10
Keskiviikko	20	19	12	11	8	7
Torstai	10	10	14	12	22	18
Perjantai	25	22	24	20	24	22
Yhteensä	**94**	**73**	**86**	**71**	**74**	**61**

10. Tarkastellaan seuraavaksi Pyöräilybarometria. Muuta muuttujan `Aq3` ”En osaa sanoa”-vastaukset puuttyviksi arvoiksi. Tutki sen jälkeen riippumattomuustestin avulla riippuuko tyytyväisyys Helsinkiin pyöräilykaupunkina (sarake Aq3)

    a) Sukupuolesta (sarake AA)?

    b) Ikäryhmästä (sarake Aikalk1)?

    c) Asuinalueesta (sarake Aalue2)?

    d) Tulotasosta (sarake AH)?