805305A JOHDATUS REGRESSIO- JA VARIANSSIANALYYSIIN, sl 2022
Harjoitus 4, viikko 38: mikroluokkatehtävät

1. Jatketaan kotitehtävissä jo käsitellyn koeaineiston analysointia käyttäen R:n työkaluja. Jos teit harjoituksen 3 R-tehtävät kokonaisuudessaan, voit hyödyntää ko. harjoituksen R-scriptitiedostoa tämän harjoituksen alkuosassa, jossa aineiston keskeisimmistä muuttujista lasketaan kuvailevien tunnuslukujen arvoja. Kokeen tarkoituksena oli siis arvioida kevyen liikuntasuorituksen vaikutusta sykkeeseen. Tarkempi kuvaus koejärjestelyistä on esitetty harjoituksen 3 tehtäväpaperin liitteessä.

Tiedosto `sykkeet2015.txt` sisältää valittujen muuttujien arvot niiltä kokeeseen osallistuneilta koehenkilöiltä ($n = 36$), joilla sykemittaus käytetyllä menetelmällä onnistui. (Kopioi tiedosto tarvittaessa Moodlesta omalle koneellesi.)

(a) Lue aineisto R:n datakehikoksi ja tutki sen rakennetta.

```
syke <- read.table("sykkeet2015.txt", header=T)
str(syke)
```

(b) Muuttuja `ryhma` on koodattu siten, että 1 = "vertailuryhmä" ja 2 = "koeryhmä". Muuta tämä muuttuja laadulliseksi tekijäksi ja anna sen tasoille selväkieliset nimet. Sen jälkeen tulosta kaikkien muuttujien suorat jakaumat ja kiinnitä datakehikko.

```
> syke$ryhma <- factor(syke$ryhma, labels = c(' vertailu', ' koe') )
> summary(syke)
> attach(syke)
```

(c) Käyttäen hyväksi skriptitiedoston `Esanfunktiot.R` sisältämää funktiota `tunnus.taulu()` laske ja tulosta tämän viikon kotitehtävässä 3 käytettävän vastemuuttujan (`sykemuutos`) ryhmäkohtaiset lukumäärät, keskiarvot, hajonnat ja varianssit. Ennen tätä lue ao. tiedosto sisään komennolla `source()`. (Kopioi tarvittaessa tiedosto `Esanfunktiot.R` Moodlesta koneellesi.)

```
> source("Esanfunktiot.R")
> tunnus.taulu(sykemuutos, ryhma, 2)
```

Totea, että luvut ovat samat kuin kotitehtävässä 3 annetut.

(d) Piirrä loppusykkeen jakauman pistekuvio ja laatikkokuvio päällekkäin kuten tehtäväpaperissa tehtävän 3 kohdalla on tehty.

```
> library(beeswarm)
> par(mfrow=c(1,1))
> beeswarm( sykemuutos ~ ryhma, horizontal = T)
> boxplot( sykemuutos ~ ryhma, horizontal = T, add = T)
```

2. Loppusykkeen odotusarvojen erotusta koskeva päättely.

(a) Laske vasteen keskiarvot kummassakin ryhmässä ja täydennä edellä laatimaasi kuviota sijoittamalla laatikoiden sisään isot mustat pisteet kuvaamaan ryhmäkeskiarvoja.

```
> ls.karvot <- tapply(sykemuutos, ryhma, mean)
> ls.karvot
> points( ls.karvot, c(1,2), pch = 16, cex = 1.5)
```

(b) Käytä R:n funktiota `t.test()` toteuttamaan kotitehtävässä 3. (c) pyydetyt laskelmat (T-suure, P-arvo, 95% luottamusväli) ja vertaa tuloksia kotitehtävän ratkaisuihin. Huomaa argumentin `var.equal` arvo vakiovarianssioletuksen mukaan ottamiseksi.

```
> t.test(sykemuutos ~ ryhma, var.equal=TRUE)
```

Moodle 1

(c) Toteuta sama analyysi kuin kohdassa (b) käyttäen nyt funktiota `lm()`, joka sovittaa lineaarisia malleja normaalijakautuneeksi oletetulle vastemuuttujalle. Tarkastele tämän ajon tuloksia ja vertaa niitä edellisen kohdan vastaaviin.

```
> lm1 <- lm(sykemuutos ~ ryhma)
> summary(lm1)
> confint(lm1)
```

Moodle 2

Mm. odotusarvojen vertailuparametriin $\delta = \alpha_2 - \alpha_1$ liittyvässä tilastollisessa päättelyssä tarvitaan käytetyn estimaattorin keskivirhettä, jonka laskennassa on keskeisessä roolissa jäännöskeskihajonta S. Jäännöskeskihajonta löytyy tulostuksen "`coefficients`"-taulun alapuolelta kohdasta `Residual standard error`. Minkä arvon jäännöskeskihajonta saa tässä mallissa? Moodle 3

Moodle 4

(d) Funktiota `lm()` käytettäessä saadaan sovitetusta mallioliosta tulostetuksi myös vastaava ANOVA-taulu sekä diagnostisia kuvioita, joiden pohjalta voi arvioida havaintojen sopusointua mallioletusten kanssa.

```
> anova(lm1)
> par(mfrow=c(1,2))
> plot(lm1, 1:2)
```

Moodle 5

Mitä edellä piirretyt diagnostiikkakuviot kertovat mallioletusten realistisuudesta? Entä miten tulostuksen ANOVA-taulun perusteella saadaan määriteltyä vasteen kokonaisvaihtelun määrää kuvaavan neliösumman SSY arvo? Tarkista päätelmäsi laskemalla neliösumman SSY arvo kaavalla $(n-1)S_Y$

```
> (length(sykemuutos)-1) * var(sykemuutos)
```

Moodle 6

Etsi vielä ANOVA-taulusta jäännösvarianssin S^2 arvo ja ota siitä neliöjuuri funktiolla `sqrt()`. Vertaa saamaasi tulosta edellä tulostuksesta poimitun jäännöskeskihajonnan arvoon 6.486. Mitä huomaat?

(e) Yksi kriittinen mallioletus, johon em. analyysit pohjautuvat, koskee virhevarianssin homoskedastisuutta (vakiovarianssioletus). Seuraavassa toteutetaan funktion `t.test()` avulla Aspinin–Welchin–Satterthwaiten approksimaatiolla (ks. luentomonisteen alaluku 4.5) vaihtoehtoinen analyysi, jossa sallitaan virhevarianssien olla erisuuruiset eri ryhmissä, eli asetetaan `var.equal=FALSE`.

```
> t.test( sykemuutos ~ ryhma, var.equal=F)
```

Moodle 7

Vertaa tuloksia kohdassa (b) saamiisi. Ovatko olennaisesti erilaiset ja vaikuttavatko päätelmiisi?

Moodle 8

(f) Tämän tehtävän lopuksi irrota datakehikko.

```
> detach(syke)
```

3. Eräässä matematiikan didaktiikan tutkimushankkeessa selvitettiin kokeellisesti mm. sitä, kuinka erityyppisen palautteen anto vaikuttaa matematiikan oppimistuloksiin alakoululaisilla. Saman luokka-asteen oppilaat jaettiin satunnaistamalla 4 eri ryhmään, joille kullekin opetettiin yhden viikon ajan päivittäisillä oppitunneilla sama oppisisältö, ja oppilaita pyydettiin näillä tunneilla myös itse ratkomaan annettuja harjoitustehtäviä. Ryhmät (muuttuja `ryhma`) ja niille annetut "käsittelyt" olivat:

A: Oppilaille ei annettu mitään palautetta oppituntien päätteeksi
B: Kuten ryhmä A
C: Koko ryhmän suoriutumista harjoitustehtävistä kehuttiin oppitunnin päätteeksi joka päivä
D: Koko ryhmälle annettiin oppitunnin päätteeksi lähinnä negatiivista, havaittuihin virheisiin painottuvaa palautetta.

Seuraavan viikon maanantaina kaikille ryhmille järjestettiin yhteinen testi, jolla mitattiin edellisen viikon aikana opetettujen asioiden osaamista. Vasteena on tästä testistä saatu pistemäärä (muuttuja `pisteet`). Oppilaskohtaiset tulokset löytyvät tiedostosta `mat-koe.txt`. Kopioi kyseinen tiedosto tietokoneellesi seuraavia analyysejä varten.

Analysoidaan aineistoa 1-suuntaisen varianssianalyysin mallilla käyttäen sekä tyypin (a) että tyypin (b) parametrointia (ks. luentomonisteen alaluku 4.1).

(a) Lue aineisto R:n datakehikoksi, tutki sen rakennetta, suoria jakaumia ja kiinnitä.

```
> mat <- read.table("mat-koe.txt", header=T)
> str(mat)
> summary(mat)
> attach(mat)
```

Moodle 9

(b) Piirrä havainnot ryhmittäin pistekuvioon siten, että ryhmä on x-akselilla ja vastemuuttuja y-akselilla

```
> par(mfrow=c(1,1))
> beeswarm(pisteet ~ ryhma, method='center')
```

Pysähdy hetkeksi tarkastelemaan kuviota. Onko ryhmien välillä eroja? Voiko kuvasta päätellä mitään mallin virhetermejä koskevien tavanomaisten oletusten uskottavuudesta? Moodle 10

(c) Laske ja tulosta vasteen ryhmäkohtaiset tunnusluvut funktiolla `tunnus.taulu()`

```
> tunnus.taulu(pisteet, ryhma, 2)
```

Mitä havaintoja teet ryhmien välisistä eroista? Moodle 11 Aiheuttavatko empiiristen hajontojen ja varianssien vaihtelut ryhmien välillä huolta homoskedastisuusoletuksen realistisuudesta?

4. Mallin sovittaminen ja parametrien estimointi:

(a) Sovita 1-suuntaisen varianssianalyysin malli tyypin (a) parametroinnilla käyttäen funktiota `lm()` ja sijoittaen tulokset objektiin `m.a`. Tulosta parametrien estimaatit, keskivirheet ja luottamusvälit tiiviimmin kuin aiemmin hyödyntämällä funktiota `cbind()`, joka "sitoo" tulostettavia sarakkeita yhteen.

```
> m.a <- lm( pisteet ~ ryhma - 1)
> round( cbind( summary(m.a)$coef, confint(m.a) ), 2)
```

Tarkastele tulostusta ja vertaa tehtävässä 3 laskettuihin ryhmäkeskiarvoihin. Mitä havaintoja ja päätelmiä teet? Moodle 12

(b) Sovita seuraavaksi periaattessa sama malli mutta noudattaen tyypin (b) parametrointia (funktion `lm()` oletusarvo). Tulosta parametrien estimaatit, keskivirheet ym. kuten edellä.

```
> m.b <- lm( pisteet ~ ryhma)
> round( cbind( summary(m.b)$coef, confint(m.b) ), 2)
```

Vertaa näitä tuloksia edellisen kohdan tuloksiin. Miten tulkitset estimaatteja tässä parametroinnissa? Moodle 13 Mitä päätelmiä teet erityyppisen palautteen antamisen vaikutuksista?

Moodle 14

5. Sovitteet, jäännöstermit ja Anova-taulu.

(a) Laske kullekin oppilaalle vastemuuttujan sovitteiden $\widehat{y}_{ki}$ ja jäännöstermien $e_{ki} = y_{ki} - \widehat{y}_{ki}$ havaitut arvot. Sovitteet saat funktiolla `fitted.values()`, jonka argumenttina on kumpi hyvänsä edellisessä tehtävässä sovitettu malliolio, eli joko `m.a` tai `m.b`. Sijoita sovitteet vektoriin `yhat`. Laske ja sijoita jäännöstermit vektoriin `res` joko tyyliin `res <- pisteet - yhat` tai suoraan mallioliosta `res <- residuals(m.b)`. Tulosta havaitut vasteen arvot, sovitteet ja jäännöstermit rinnakkain. Merkitse myös havaintojen pistekuvioon ryhmäkohtaiset sovitteet (eli ryhmäkeskiarvot) mustalla ympyrällä.

```
> yhat <- fitted.values(m.b)
> res <- residuals(m.b)
> data.frame(ryhma, pisteet, yhat = round(yhat, 2), res = round(res,2))
> points( as.numeric(ryhma) , yhat, pch = 16, cex = 1.5)
```

Moodle 15

(b) Tulosta mallin ANOVA-taulu.

```
> anova(m.b)
```

Mitä päätelmiä voit tämän taulukon pohjalta tehdä ryhmien välisistä systemaattisista eroista? Moodle 16 Miten saat selville kokonaisneliösumman SSY arvon? Moodle 17

(c) Mallin oletusten realistisuuden graafista tarkastelua varten piirrä jäännöstermit sovitteita vastaan sekä jäännöstermien QQ-kuvio:

```
> par(mfrow=c(1,2)) #  kaksi kuvaa rinnakkain
> plot(m.b, which=1:2)
```

Miten tulkitset näitä kuvioita?

6. Kontrastien estimointi.

(a) Olisiko niin, että mikä tahansa palaute olisi yhtä hyvä kuin ei palautetta lainkaan? Tällöin voisi yhdistää ryhmät C ja D ja verrata testipisteiden odotusarvoja ao. yhdistetyn ryhmän C&D sekä yhdistetyn ryhmän A&B välillä, ottaen huomioon että ryhmät A ja B saivat alun perin samanlaisen "käsittelyn".

Tällaisen kontrastin kerroinvektori $(c_1, \ldots, c_4)$ on muotoa $(-0.5, -0.5, 0.5, 0.5)$. Lasketaan nyt ao. kontrastin piste-estimaatti, keskivirhe, T-suure, P-arvo sekä 95% luottamusväli käyttäen paketin `gmodels` sisältämää funktiota `fit.contrast()` seuraavasti (asenna tarvittaessa lisäpaketti `gmodels` R:ään):

```
> library(gmodels)
> fit.contrast(m.b, ryhma, coeff = c(-1, -1, 1, 1)/2, conf.int=0.95 )
```

Vertaa saamaasi tulosta, erityisesti P-arvoa ja luottamusväliä mallin `m.b` sovituksesta saatuihin vastaaviin tuloksiin, jotka koskevat kertoimia `ryhmaC` ja `ryhmaD`. Moodle 18

(b) Tarkastetaan silti, millainen näyttö on siitä, tuottaako positiivinen palaute keskimäärin paremmat testipistemäärät kuin negatiivinen. Sovella funktiota `fit.contrast()` kuten edellä, mutta nyt kontrastin kerroinvektori on muotoa $(0, 0, +1, -1)$ Mitä päättelet? Moodle 19

(c) Palautteen laadulla on siis väliä. Vaan millainen on lopulta kontrasti negatiivisen palautteen ja "ei palautetta lainkaan" -käsittelyn välillä? Kontrastin kerroinvektori on siten $(-0.5, -0.5, 0, 1)$. Moodle 20