805305A JOHDATUS REGRESSIO- JA VARIANSSIANALYYSIIN, sl 2022
Harjoitus 3, viikko 37: mikroluokkatehtävät

1. Aloitetaan harjoituksen R-osuus analysoimalla kotitehtävissä esiteltyä komponentin elinaika -aineistoa käyttämällä R:n työkaluja.

(a) Talleta havaintoarvot vektoriin `elinaika`:

```
> elinaika <- c(4, 5, 10, 11, 20, 29, 35, 40, 66, 70)
```

(b) Piirrä havainnoista pistekuvio funktiolla `beeswarm()`; asettaen `horizontal=TRUE`. Lataa sitä ennen paketti beeswarm käyttöösi komennolla `library(beeswarm)`.

```
> library(beeswarm)
> beeswarm(elinaika,  horizontal=TRUE)
```

(c) Piirrä edellinen kuva perus-R:n mukana tulevalla funktiolla `stripchart()`:

```
> stripchart(elinaika)
```

Tutki ao. funktion help-sivua komennolla `?stripchart` ja etsi minkä argumentin avulla kuva saadaan piirrettyä oletusarvon sijaan pystysuoraan? Moodle 1

(d) Laske komponenttien elinaikojen sijaintia ja hajontaa kuvaavia tunnuslukuja ja totea, että keskiarvo ja keskihajonta vastaavat kotitehtävän 2 tehtävänannossa ilmoitettuja arvoja

```
> summary(elinaika) ; round(sd(elinaika), 1)
```

(e) Suorita nollahypoteesin H_0: $\mu = 40$ merkitsevyystestaus ja laske parametrille μ 95 % luottamusväli. Toteuta pyydetty analyysi funktiolla `t.test()`. Katso tarvittaessa apua tarvittavan komennon muotoiluun esimerkiksi kyseisen komennon help-sivuilta (`?t.test`) tai mikroharjoituksen 2 tehtävästä 1 (i).

Vertaa saamaasi tulostusta kotitehtävän 2 kohdissa (b) ja (c) saatuihin tuloksiin. Moodle 2

2. Varianssi- ja regressioanalyyseissä jatkuvalla vastemuuttujalla on joko yksi tai useampi selittäjä ja tällöin käytämme tarvittavien analyysien suorittamiseen usein funktiota `lm()` (*lm = linear model*). Kokeillaan seuraavaksi `lm()`-funktion toimintaa asetelmassa, jossa vasteella ei ole yhtään selittäjää (vrt. edellinen tehtävä), jolloin analysoitava malli on muotoa $Y_i = \mu + \epsilon_i \;\; (i = 1, ..., 10)$.

(a) Funktiossa `lm()` malli määritellään muodossa "*vaste* $\sim$ *selittäjä(t)*" ja mallituksen tulokset kannattaa yleensä tallettaa erilliseen malliobjektiin. Mallitetaan seuraavaksi vastemuuttujaa mallilla, jossa ainoana selittäjänä on vakiotermi (tässä tapauksessa vasteen odotusarvo).

```
> malli <- lm(elinaika ~ 1)
```

(b) Muodostetusta malliobjektista voidaan poimia estimoidun mallin mukaiset Y_i:n sovitetut arvot ($\widehat{y}_i$) ja jäännöstermit eli residuaalit ($e_i = y_i - \widehat{y}_i$) kaikille havaintoyksiköille funktioilla `fitted()` ja `resid()`.

```
> sovitteet <- fitted(malli) ; sovitteet
```
Moodle 3

```
> residuaalit <- resid(malli) ; residuaalit
```
Moodle 4

Tulosta vasteen alkuperäiset arvot, sovitteet ja residuaalit datakehikkona.

```
> data.frame(elinaika, sovitteet, residuaalit)
```

Vertaa saatua tulostusta kotitehtävän 3 (b) vastauksiin.

(c) Laske residuaalien keskiarvo `mean()` ja keskihajonta `sd()`. Vertaa saatuja arvoja vastemuuttujan vastaavien tunnuslukujen havaittuihin arvoihin.

(d) Laske sovitteiden keskiarvo ja keskihajonta. Vertaa saatuja arvoja jälleen vastemuuttujan vastaavien tunnuslukujen havaittuihin arvoihin. Moodle 5

3. Jatkoa edelliseen tehtävään.

(a) Tutkitaan seuraavaksi mitä muodostetusta malliobjektista saadaan tulostettua funktiolla `summary()`.

```
> summary(malli)
```

Mallin systemaattisen osan määrittelevän parametrin μ piste-estimaattorina toimii vastemuuttujan otoskeskiarvo eli $\widehat{\mu} = \bar{Y}$, jonka havaittua arvo on 29. Estimaattorin keskihajonta eli keskivirhe saadaan laskettua kaavalla $S_Y/\sqrt{n}$.

```
> sd(elinaika)/sqrt(length(elinaika))
```

Etsi edellä lasketun estimaattorin keskivirheen arvo komennolla `summary(malli)` aikaansaadusta tulostuksesta. Moodle 6 `summary()`-funktion tulostuksessa estimaattorin keskivirheen viereen on tulostettu testisuureen havaittu arvo ja siihen liittyvä P-arvo. Suorita seuraavaksi `elinaika`-muuttujalle merkitsevyystestaus funktiolla `t.test()`

```
> t.test(elinaika, mu=0, conf.level=0.95)
```

Vertaa saadun tulostuksen testisuureen arvoa ja P-arvoa `summary()`-funktiolla aikaansaadun tulostuksen testisuureen havaittuun arvoon ja P-arvoon. Mitä huomaat? Mitä ko. testeissä on testattu? Moodle 7 Onko kyseinen testaus käytännön kannalta ajatellen järkevä?

(b) Malliobjektista `malli` voidaan tulostaa parametrin μ 95 % luottamusväli funktiolla `confint()`.

```
> confint(malli)
```

Vertaa luottamusväliä edellisessä kohdassa `t.test()`-funktiolla saatuun luottamusväliin. Selvitä (`?confint`) millä argumentilla `confint()`-funktiossa määritellään laskettavan luottamusvälin luottamustaso ja laske odotusarvon 99 % luottamusväli. Onko valmistajan väite komponentin keskimääräisestä 40 viikon eliniästä uskottava lasketun luottamusvälin perusteella? Moodle 8

4. Jatkoa edelliseen tehtävään. Piirretään seuraavaksi vastemuuttujan arvoihin liittyvä QQ-kuvio, jonka avulla on mahdollista arvioida kuvion kohdemuuttujan normaalijakautuneisuutta. Piirrettävässä (sironta)kuviossa yksittäisen pisteen y-akselin koordinaatin määrittää vastemuuttujan havaittu arvo y_k $(k = 1, \ldots, n)$ ja x-akselin koordinaatti $z_{[k]}$ saadaan määrättyä $N(0,1)$-jakauman fraktiilipisteiden avulla. Näiden $z_{[k]}$-lukujen tulee täyttää ehto $P(Z \leq z_{[k]}) = k/(n+1)$. Piirrettävässä kuviossa yksittäinen $z_{[k]}$-luku voidaan tulkita kohdemuuttujan Y standardoituna (normeerattuna) versiona eli $z_{[k]} = \frac{Y_{[k]}-\mu}{\sigma}$.

(a) Lasketaan ensin x-akselille tarvittavat $z_{[k]}$-fraktiilit.

```
> n <- length(elinaika) ; n
> i <- 1:n ; i
> nu <- i/(n+1) ; nu
> z.k <- qnorm(nu) ; z.k
```

Moodle 9

Tarkista seuraavaksi, vastaako vektorin `z` ensimmäisen (ja samalla pienimmän) alkion kohdalla vaadittu ehto $P(Z \leq z_{[1]}) = 1/(n+1)$.

```
> qnorm(1/(n+1))
> z.k[1]
```

QQ-kuvio voidaan piirtää nyt tavanomaisella piirtofunktiolla `plot()`.

```
> plot(z.k, elinaika, main="QQ-kuvio")
```

Jos vastemuuttuja jakautuu kohdepopulaatiossa normaalijakauman kaltaisesti, voidaan pisteiden odottaa sijoittuvan QQ-kuviossa ainakin likimain samalle nousevalle suoralle. Onko mielestäsi piirretyn kuvion perusteella oletus komponenttien eliniän normaalijakautuneisuudesta realistinen?

(b) Vastaavantyyppinen kuvio voidaan piirtää myös esimerkiksi funktiolla `qqnorm()`, jonka jälkeen piirrettyyn kuvaan on helppo lisätä edellä mainittu referenssisuora funktiolla `qqline()`.

```
> qqnorm(elinaika)
> qqline(elinaika)
```

(c) Piirretään vielä samaan kuvaan alekkain QQ-kuviot alkuperäiselle vasteelle ja mallituksesta saataville residuaaleille eli jäännöstermeille.

```
> par(mfrow=c(2,1))
> qqnorm(elinaika) ; qqline(elinaika)
> qqnorm(residuaalit) ; qqline(residuaalit)
```

Vertaile edellä piirrettyjä kahta kuvaa toisiinsa. Poikkeavatko ne toisistaan? Moodle 10

5. Analysoidaan seuraavaksi kotitehtävässä 1 esitellyn kokeen tuloksia. Kokeessa pyrittiin arvioimaan kevyen liikuntasuorituksen vaikutusta sykkeeseen. Tarkempi kuvaus koejärjestelyistä on kotitehtävän liitteessä.

Tiedosto `sykkeet2015.txt` sisältää valittujen muuttujien arvot niiltä kokeeseen osallistuneilta koehenkilöiltä, joilla sykemittaus käytetyllä menetelmällä onnistui. Kopio tiedosto Moodlesta tietokoneesi tämän kurssin työhakemistoon, ja lue se R: muistiin alla esitettyyn tapaan funktiolla `read.table()`. Muista asettaa R-istuntosi työhakemisto seuraamalla RGui-ikkunan vasemmasta ylänurkasta lähtien valikkopolkua `File - Change dir ...`

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

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

Moodle 11

(b) Muuttuja `ryhma` on koodattu tiedostoon seuraavasti: 1 = "vertailuryhmä", 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 vastemuuttujan ryhmäkohtaiset lukumäärät, keskiarvot, hajonnat ja varianssit. Ennen tätä lue ao. tiedosto sisään komennolla `source()` ja tutki, miten funktio `tunnus.taulu` on ohjelmoitu, mitkä ovat sen syötteet ja tulosteet. Kyseistä skriptitiedostoa käytettiin jo aiemmin R-harjoituksessa 2, mutta jos et ole kopioinut kyseistä tiedostoa (`Esanfunktiot.R`) aiemmin Moodlesta, kopioi se ennen seuraavia komentoja.

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

Moodle 12

(d) Piirretään seuraavaksi loppusykkeen jakauman pistekuvio ja laatikko-janakuvio päällekkäin.

```
> par(mfrow=c(1,1))
> beeswarm( loppusyke ~ ryhma, horizontal = TRUE)
> boxplot( loppusyke ~ ryhma, horizontal = TRUE, add = TRUE)
```

Moodle 13

6. Loppusykkeen odotusarvojen erotusta koskeva päättely.

(a) Laske vasteen keskiarvot kummassakin ryhmässä sekä keskiarvojen erotus kahdessa eri ”suunnassa”. Täydennä myös edellisessä tehtävässä laatimaasi kuviota sijoittamalla laatikoiden sisään isot mustat pisteet kuvaamaan ryhmäkeskiarvoja.

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

Moodle 14

(b) Käytä R:n funktiota `t.test()` toteuttamaan ryhmien välisen odotusarvojen vertailun.

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

Moodle 15

Huomaa argumentin `var.equal` arvo, jolla asetetaan oletus vertailtavien ryhmien varianssien yhtäsuuruudesta. Missä suunnassa `t.test()` raportoi keskiarvojen vertailun eli kumpi ryhmistä toimii vertailuolosuhteena? Mikä on kyseisen testin nollahypoteesina ja mikä johtopäätäs laskelmien perusteella voidaan tehdä?

(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. Mitä yhtäläisyyksiä ja mitä eroja havaitset tulostusten välillä?

```
> lm1 <- lm(loppusyke ~ ryhma)
> summary(lm1)
```

Tarkastele tämän ajon tuloksia ja vertaa niitä edellisen kohdan vastaaviin. Mitä yhtäläisyyksiä ja mitä eroja havaitset tulostusten välillä? Löytyykö tulostuksista esimerkiksi samoja testisuureiden havaittuja arvoja ja P-arvoja? Moodle 16

(d) Tulostetaan seuraavaksi muodostetusta malliobjektista 95 % luottamusvälit.

```
> confint(lm1)
```

Vertaa jälleen saatua tulostusta (b)-kohdassa saatuun tulostukseen. Onko tuloksissa nähtävissä jotain yhteisiä elementtejä? Moodle 17

(e) Poimitaan muodostetusta malliobjektista `lm1` seuraavaksi sovitteet ja residuaalit ja tulostetaan ne datakehikkona yhdessä vasteen eli loppusykkeen havaittujen arvojen kanssa.

```
> sov <- fitted(lm1); sov
> res <- resid(lm1)
> data.frame(loppusyke, sov, res)
```

Moodle 18

Moodle 19

(f) Funktiota `lm()` käytettäessä saadaan muodostettavasta malliobjektista tulostetuksi myös mm. erilaisia diagnostisia kuvioita, joiden pohjalta voi arvioida havaintojen sopusointua mallioletusten kanssa. Tulostetaan tämän harjoituksen lopuksi kaksi diagnostiikkakuvaa.

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

Tutki millaisia kuvia ruudulle tulostuu. Vasemmanpuoleisen kuvion perusteella voidaan arvioida vakiovarianssioletuksen realistisuutta ja oikeanpuoleinen kuvio on jo edellä esitelty QQ-kuvio, jonka avulla voidaan arvioida virhetermeihin liittyvän normaalijakaumaoletuksen realistisuutta.

Moodle 20