Tilastotieteen peruskurssi **5 op** Tentti 26.3.2021

Tentissä on neljä kysymystä, ja jokaisesta maksimipistemäärä on 12 pistettä. Kaikki laskut voi tehdä ohjelmistolla tai käsin. Jos lisäät vastauksiisi R-ohjelmalla tehdessäsi käyttämäsi koodit ja käsin tehdessä välivaiheet, voit näiden perusteella saada osapisteitä vaikka lopullinen vastaus on väärin. Tehtäväpaperin R-koodeista poimittuja arvoja ei tarvitse laskea uudestaan mutta niiden käyttö tulee perustella. Ellei tehtävänannossa muuta sanota, käytä testauksessa merkitsevyystasona $\alpha = 0.05$. Menestystä tenttiin!

1. Mantukimalaiskoiraat voidaan luokitella karvapeitteen värin mukaan tummiin, välimuotoa oleviin ja vaaleisiin yksilöihin. Toisena luokittelijana on kimalaisen koko; suuri ja pieni. Onko kimalaisen koolla ja värillä yhteyttä? Aseta tilastolliset hypoteesit, valitse sopiva menetelmä, kirjaa valitsemasi menetelmän vaatimat oletukset ja tee menetelmän avulla tarvittavat päätelmät.

väri	koko	lkm
tumma	pieni	32
tumma	suuri	12
vaalea	pieni	6
vaalea	suuri	9
välimuoto	pieni	14
välimuoto	suuri	22

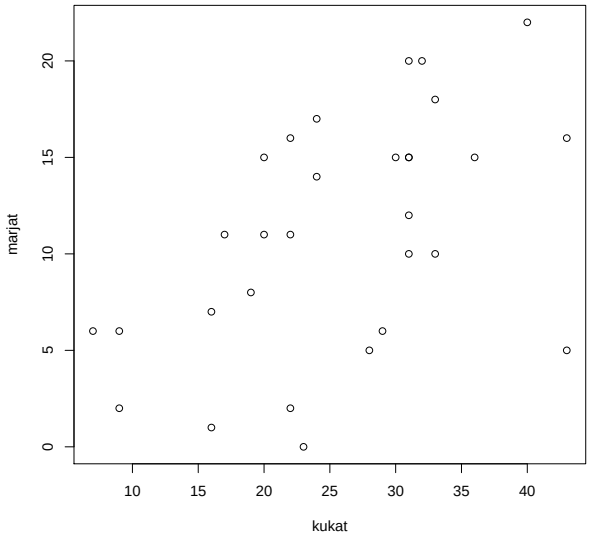
2. Mustikka kukkii aikaisin keväällä. Kukista muodostuu raakileita kesäkuussa ja raakileet kypsyvät heinä-elokuussa. Jos mustikan kukkien pölytys onnistuu täydellisesti, eivätkä esimerkiksi tuholaiset tai pakkasyöt tuhoa kukkia, jokaisesta kukasta voi tulla yksi marja. Tutkija halusi selvittää kukkien ja marjojen määrän välistä yhteyttä Jaamankankaan männikössä vuonna 2020. Keväällä kukinnan aikaan perustettiin 30 kappaletta 0.25 m^2 koealaa, joiden paikat arvottiin satunnaisesti, ja kultakin koealalta laskettiin mustikan kukintojen määrä. Juhannuksen aikaan, ennen marjojen kypsymistä samoilta koealoilta kirjattiin ylös mustikan raakileiden lukumäärä. Kypsiä mustikoita ei tutkittu, koska poiminta vaikuttaa niiden määrään. Alla on aineistosta piirretty sirontakuvio sekä tulosteita aineistoon sovitetuista regressiomalleista.

(a) Kirjoita aineistoon sovitetun mallin perusteella yhtälö, joka kuvaa miten marjojen määrän odotusarvo (populaatiokeskiarvo) riippuu kukkien määrästä.

(b) Tulkitse mallin regressiokertoimen estimaatti. Mikä on mallin vakiotermin tulkinta ja onko se mielekäs?

(c) Paikallisen sanomalehden marjojen kukintaan liittyvän artikkelin keskustelupalstalla nimimerkki *marjastaja* kirjoitti: “Ihan sama oliko kukkia paljon tai vähän,

saman verran niitä marjoja tulee kuitenkin. Riittävästi kukkia on joka paikassa ja jos niitä on paljon, niin pieni osa vain muodostuu marjaksi". Tutki onko aineistossa todisteita *marjastajan* väitettä vastaan.




```
> summary(malli1)

Call:
lm(formula = marjat ~ kukat, data = jaama)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5260     2.8575   0.884  0.38423
kukat         0.3306     0.1045   3.164  0.00373 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.293 on 28 degrees of freedom
Multiple R-squared:  0.2633,Adjusted R-squared:  0.237
F-statistic: 10.01 on 1 and 28 DF,  p-value: 0.003733

> summary(malli2)

Call:
lm(formula = kukat ~ marjat, data = jaama)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9451     3.1570   5.367 1.02e-05 ***
marjat        0.7965     0.2518   3.164  0.00373 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.216 on 28 degrees of freedom
Multiple R-squared:  0.2633,Adjusted R-squared:  0.237
F-statistic: 10.01 on 1 and 28 DF,  p-value: 0.003733
```

3. Tutkija haluaa selvittää auton renkaiden vaikutusta polttoaineen kulutukseen. Testissä oli renkaita neljää eri tyyppiä, numeroidut ryhmät 1, 2, 3 ja 4. Kutakin rengastyyppiä

oli kymmenen rengassettiä ja kutakin ajettiin 1000 km samaa testireittiä käyttäen. Testiajolta mitattiin polttoaineenkulutus litraa/100 km. Testaa sopivaa menetelmää käyttäen, onko renkaan tyypillä vaikutusta polttoaineen kulutukseen ja mitkä rengastyypit eroavat toisistaan kulutuksen osalta.

```
aggregate(datdf$kulutus, by = list(datdf$tyyppi),FUN = mean)
Group.1     x
1       1 6.701
2       2 7.565
3       3 5.185
4       4 6.582

> library(car)
> leveneTest(datdf$kulutus, datdf$tyyppi)
datdf$tyyppi coerced to factor.Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.3701  0.775
      36

> m1 <- aov(kulutus ~ factor(tyyppi), data=datdf)
> anova(m1)
Analysis of Variance Table

Response: kulutus
               Df Sum Sq Mean Sq F value    Pr(>F)
factor(tyyppi)  3 29.103  9.7010  18.387 2.109e-07 ***
Residuals      36 18.994  0.5276
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m2 <- lm(kulutus ~ tyyppi, data=datdf)
> anova(m2)
Analysis of Variance Table

Response: kulutus
          Df Sum Sq Mean Sq F value  Pr(>F)
tyyppi     1  3.746  3.7456  3.2092 0.08119 .
Residuals 38 44.352  1.1671
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m3 <- lm(kulutus ~ factor(tyyppi), data=datdf)
> anova(m3)
Analysis of Variance Table

Response: kulutus
               Df Sum Sq Mean Sq F value    Pr(>F)
factor(tyyppi)  3 29.103  9.7010  18.387 2.109e-07 ***
Residuals      36 18.994  0.5276
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pairwise.t.test(datdf$kulutus,datdf$tyyppi,p.adj="none")

Pairwise comparisons using t tests with pooled SD

data:  datdf$kulutus and datdf$tyyppi

  1       2       3
2 0.01160 -       -
3 4.1e-05 1.2e-08 -
4 0.71626 0.00455 0.00012

P value adjustment method: none
```

```
> pairwise.t.test(datdf$kulutus,datdf$tyyppi,p.adj="bonferroni")

Pairwise comparisons using t tests with pooled SD

data:  datdf$kulutus and datdf$tyyppi

  1       2       3
2 0.06960 -       -
3 0.00025 7.4e-08 -
4 1.00000 0.02733 0.00075

P value adjustment method: bonferroni
```

4. Valitse oikea vaihtoehto. Kussakin kysymyksessä **täsmälleen yksi annetuista vaihtoehdosta** on oikein. Kysymyksen pisteytys on seuraava:
Oikea vastaus = 2 pistettä,
väärä vastaus = -1 pistettä,
ei vastausta = 0 pistettä.
Huomaa, että tehtävän kokonaispistemäärä voi olla myös negatiivinen. Jos kokonaispistemäärä on negatiivinen, se vähentää muista kysymyksistä saatuja pisteitä.

(a) Tarkastellaan satunnaismuuttjia X ja Y, joiden yhteisjakauman pistetodennäköisyysfunktio on esitetty alla. Mikä seuraavista väittämistä **ei pidä paikkaansa**?

		X		
		1	2	3
Y	1	0.2	0.4	0.0666...
	2	0.1	0.2	0.0333...

i. Muuttujat X ja Y ovat riippumattomia.

ii. Muuttujat X ja Y ovat korreloimattomia.

iii. Muuttujien X ja Y yhteisjakauma sisältää enemmän informaatiota kuin muuttujien X ja Y marginaalijakaumat.

(b) Satunnaismuuttujille X ja Y pätee $E(X) = 0$, $E(Y) = 0$, $var(X) = 1$, $var(Y) = 4$ ja $cor(X, Y) = 0.2$. Mikä seuraavista väittämistä on tosi.

i. Kovarianssia $cov(X, Y)$ ei tunneta.

ii. $E(XY) = 0.2 \times \sqrt{1} \times \sqrt{4} = 0.4$

iii. $cov(X, Y) = 0.2 \times 1 \times 4 = 0.8$

(c) Ao. tuloste saatiin sovittamalla kaksisuuntaisen varianssianalyysin malli aineistoon, jossa vastemuuttuja `y` on suhdeasteikollinen, muuttuja `x1` voi saada arvon `A`, `B` tai `C` ja muuttuja `x2` arvon `YES` tai `NO`.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.00    0.28931  17.420  < 2e-16 ***
x1B             3.00    0.40914   7.278 1.46e-09 ***
x1C             2.50    0.40914   5.970 1.90e-07 ***
x2YES          -2.00    0.40914  -5.377 1.66e-06 ***
x1B:x2YES       0.00    0.57861   0.126    0.900
x1C:x2YES       1.00    0.57861   1.569    0.122
```

Kun `x1=C` ja `x2=YES`, muuttujan y populaariokeskiarvon estimaatti on

i. 6.5

ii. 6

iii. 1

(d) Vuonna 2019 Itä-Suomen yliopiston henkilökunta teki yhteensä reilut 3000 ulkomaanmatkaa. Matkailun aiheutamien hiilipäästöjen arvioimista varten yliopiston matkalaskujärjestelmästä poimittiin 20 ulkomaanmatkaa satunnaisesti ja matkatositteiden perusteella laskettiin kunkin otokseen poimitun matkan hiilidioksidipäästö. Otokseen poimittujen matkojen keskimääräiseksi hiilidioksidipäästöksi saatiin 960 kg ja keskihajonnaksi 1116 kg. Otoksen perusteella ulkomaan matkan hiilidioksidipäästön populaatiokeskiarvon (odotusarvon) 95%:n luottamusväli on

i. 843 – 1077 kg

ii. 437 – 1482 kg

iii. -156 – 2076 kg

(e) Puut valmistautuvat talven mm. väkevöittämällä solunesteitä talven lähestyessä. Siksi puut kestävät koviakin pakkasia talvisaikaan, kun taas pakkaset kesällä tai alkusyksyllä tappaisivat puun. Tutkija halusi selvittää, miten pakkaskestävyys kehittyy syksyn aikana. Kokeessa männyntaimia laitettiin pakastearkkuun viikoksi eri ajankohtina ja tutkittiin, selvisikö taimi pakastamiskäsittelystä. Analyysi tehtiin logistisella regressiomallilla, jossa selittäjänä oli pakastamisen ajankohta (päivää kokeen alusta; koe aloitettiin syyskuun alussa ja sitä jatkettiin 50 päivän ajan) ja vastemuutujana oli binaarinen muuttuja, joka kertoo kuoliko taimi kokeen seurauksena vai ei (kuolema on koodattu ykkösenä). Estimoiduksi malliksi saatiin

$$\log\left(\frac{p}{1-p}\right) = 19.76 - 0.75aika\,.$$

Tarkastellaan taimien pakkaskestävyyttä syyskuun 25. päivänä, eli kun $aika = 25$.

i. Syyskuun 25. päivänä pakastettu taimi kuolee todennäköisemmin kuin säilyy hengissä.

ii. Syyskuun 25. päivänä pakastettu taimi säilyy todennäköisemmin hengissä kuin kuolee.

iii. Syyskuun 25. päivänä pakastettu taimi kuolee todennäköisyydellä 2.75.

(f) Mikä seuraavista väittämistä on totta

i. Epäparamerisia testejä tulee käyttää aina kun populaation jakauma ei ole normaalinen.

ii. Parametrisia testejä voidaan käyttää isoissa otoksissa, vaikka populaation jakauma ei olekaan normaalinen.

iii. Koska epäparametriset testit eivät tee mitään oletusta populaation jakaumasta, niitä voidaan käyttää turvallisesti kaikissa tilanteissa.